# ABSTRACT

Title of dissertation: PATTERNS AND COMPLEXITY IN BIOLOGICAL SYSTEMS: A STUDY OF SEQUENCE STRUCTURE AND ONTOLOGY-BASED NETWORKS

Kimberly Glass, Doctor of Philosophy, 2010

Dissertation directed by: Professor Michelle Girvan
Department of Physics

Biological information can be explored at many different levels, with the most basic information encoded in patterns within the DNA sequence. Through molecular level processes, these patterns are capable of controlling the states of genes, resulting in a complex network of interactions between genes. Key features of biological systems can be determined by evaluating properties of this gene regulatory network. More specifically, a network-based approach helps us to understand how the collective behavior of genes corresponds to patterns in genetic function.

We combine Chromatin-Immunoprecipitation microarray (ChIP-chip) data with genomic sequence data to determine how DNA sequence works to recruit various proteins. We quantify this information using a value termed "nmer-association." "Nmer-association" measures how strongly individual DNA sequences are associated with a protein in a given ChIP-chip experiment. We also develop the "split-motif" algorithm to study the underlying structural properties of DNA sequence independent of wet-lab data. The "split-motif" algorithm finds pairs of DNA motifs which

preferentially localize relative to one another. These pairs are primarily composed of known transcription factor binding sites and their co-occurrence is indicative of higher-order structure. This kind of structure has largely been missed in standard motif-finding algorithms despite emerging evidence of the importance of complex regulation.

In both simple and complex regulation, two genes that are connected in a regulatory fashion are likely to have shared functions. The Gene Ontology (GO) provides biologists with a controlled terminology with which to describe how genes are associated with function and how those functional terms are related to each other. We introduce a method for processing functional information in GO to produce a gene network. We find that the edges in this network are correlated with known regulatory interactions and that the strength of the functional relationship between two genes can be used as an indicator of how informationally important that link is in the regulatory network. We also investigate the network structure of gene-term annotations found in GO and use these associations to establish an alternate natural way to group the functional terms. These groups of terms are drastically different from the hierarchical structure established by the Gene Ontology and provide an alternative framework with which to describe and predict the functions of experimentally identified groups of genes.

# PATTERNS AND COMPLEXITY IN BIOLOGICAL SYSTEMS: A STUDY OF SEQUENCE STRUCTURE AND ONTOLOGY-BASED NETWORKS

by

Kimberly Glass

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2010

Advisory Committee:
Professor Michelle Girvan, Chair/Advisor
Professor Edward Ott
Professor Wolfgang Losert
Professor Doron Levy
Professor James Yorke, Dean's Representative

# Dedication

This work is dedicated to Tony D. Glass. His genuine spirit and incredible perserverance has inspired me to never give up and to keep reaching toward the impossible. Thank you Grand-dad.

# Acknowledgments

There are many people who deserve thanks in helping me get to this point in my career. Although I mention many here, others will inadvertently go unrecognized. To all of you who have who have been influential in my life, on both a personal and professional level, know that your help and support is appreciated and that I am grateful.

The first people I want to thank are my teachers, from grade-school though graduate school. I have no doubt that it is through the collection of your efforts that I am where I am today. Thank you Ms. Doherty for caring about your students, Mr. Lacour for making them laugh with all your quirks, and Mr. Frazier for seeing the best in them. Thank you Ms. Saenz for making history more than just facts, Ms. Fort for teaching real-world skills as well as English grammer and Mr. Mullen for giving respect in addition to calculus lessons. Thank you Dr. Martinich for challenging everyone to think for themselves, Dr. Moore for opening my eyes to the world of comedy and Dr. Guy for making me happy I was taking a 7 a.m. class on Laplace Transforms. For the other teachers not mentioned, and there are many who deserve recognition, know that I am forever grateful for the sacrifice you make every day to help the world to be a better place, one student at a time.

I also wish to thank a number of people from the NIH who have helped to make this thesis possible, either with advice and encouragement or with the generation of data. Thank you Julian Rozenberg, Paramita Bhattacharya and Raghu Chatterjee for giving me access to your precious hard-earned data. Thanks to Andrey

Shlyakhtenko, Vikas Rishi and Raghu Chatterjee for stimulating discussions and advice. Thank you Jeffrey Zhao for your friendship. I am also especially grateful for the help of Dr. Peter FitzGerald, whose involvement in a number of projects has not only been enjoyable but led to exciting conclusions. Finally, a big thanks to Dr. Charles Vinson, who for three years let a little physicist enter his world and play.

I also want to acknowledge the support of the University of Maryland physics department and staff. The environment you provide for your students and the positive encouragement you give them deserves commendation. A special thanks goes to Jane Hessing who has worked with me countless times to solve all those nasty bureaucratic problems. I would also like to thank several specific members of the faculty who have made my graduate career not only a success, but an enjoyable experience. Thank you Dr. Losert for introducing me to the field of systems biology, where I doubt I would have wandered of my own accord. Thank you Dr. Ott for not only being an exceptional teacher, but a thoughtful and thourough mentor. And, finally, thank you Dr. Girvan for being an exceptional advisor who was always there to help me work through all those annoying issues that come up in research.

I also want to take this opportunity to honor my ancestors. It is through your past self-sacrifice that I have been given so many wonderful opportunities in life. The successes I achieve are not only mine, but yours as well. Thanks also goes to my immediate family. Thank you Coleman and Dorothy Glass for your constant encouragement, support and love. Thank you so much for being my parents - I cannot imagine life without you. Thank you Beth Primm for being such a supportive and loving sister. If everyone in this world was as caring and thoughtful

as you, it would be a much better place. And finally, thanks to Chris Glass for your understanding and love. Your encouragement has carried me through many hard times.

And most importantly, I wish to acknowledge the love and support given to me by Travis Pittman. Even when life is hard, you encourage me to be true to myself and follow my dreams. You always know how to make me laugh when I am down and relax when I am stressed. Without you I don't know if I would be where I am today. Thank you so very much.

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| A | adenine |
| C | cytosine |
| G | guanine |
| T | thymine |
| | |
| DNA | Deoxyribonucleic Acid |
| RNA | Ribonucleic Acid |
| mRNA | messenger-RNA |
| TF | transcription factor |
| TFBS | transcription factor binding site |
| TSS | transcriptional start site |
| bp | base-pairs |
| CpG | cytosine-phosphate-guanine |
| CGI | CpG Island |
| | |
| RNAP | RNA Polymerase-II |
| H3K9me2 | Histone 3 with di-methylation of the 9th lysine |
| CREB | cAMP response element binding protein |
| C/EBP$\alpha$ | CCAAT-enhancer-binding protein alpha |
| bZip | basic leucine zipper transcription factor protein |
| | |
| NIH | National Institute of Health |
| NCI | National Cancer Institute |
| ChIP-chip | Chromatin Immunoprecipitation followed by microarray analysis (chip) |
| PCR | Polymerase Chain Reaction |
| | |
| GO | Gene Ontology |
| DAG | Directed Acyclic Graph |
| MI | Mutual Information |
| CLR | Context-Likelihood-of-Relatedness |

# Chapter 1

# Introduction

## 1.1 Complex Structure in Biological Systems

A fundamental question in biology is how the activities of thousands of different genes are coordinated in a living cell to carry out various biological processes. Each gene is controlled by other genes, which in turn are regulated by yet other genes, forming a complex, inter-connected gene regulatory network. This complex process of regulation can be understood at many different levels, from the individual mechanisms which control a gene to how that regulatory information travels on a genetic network. The recognition of these complex problems has lead to a new interdisciplinary field in biology known as "systems biology." This field represents a collaboration between formally disparate disciplines, including biology, computer science, mathematics, statistics and physics. This thesis demonstrates a physics perspective in the analysis of information contained in different levels of biological regulation.

### 1.1.1 The function of DNA sequence in biological regulation

Genes are regions of DNA which contain the information necessary for the cell to make a protein. Whether a gene is active, meaning whether that protein is made, is controlled by regions of DNA, called promoters, often physically located at

the beginning of the genes. Promoters are very short relative to the length of the genome and even compared to the typical length of a gene. Activation of a gene occurs when a set of regulatory proteins, called transcription factors (TFs) interact with each other and the promoter of that gene. In general, the set of TFs needed for activation is unique for each gene. (For more background information and a detailed description of this process see Appendix A.)

One of the primary approaches to uncovering the mechanisms by which a gene is dynamically controlled is by hunting for patterns in the DNA sequence of promoters which have the potential to interact with TFs. This is done both by analyzing laboratory data concerning TFs and also by hunting for DNA sequence patterns in promoters. The length of DNA needed to interact with a transcription factor is physically much smaller than both the size of a transcription factor and the length of a typical promoter. This means that promoter regions often contain a series of short informational patterns which together can interact with several TFs simultaneously.

Determining the critical information in promoters which distinguishes them from other areas of the genome is a current area of investigation. Another important question is how this information is used to control the behavior of genes. Both wet-lab and computational techniques are critical in understanding genetic control by transcriptional factors because they each address half of this question. Namely, laboratory data addresses the issue of how the information in a DNA sequence can control the behavior of a gene, whereas computational methods which hunt for statistically over-represented patterns give a sense of how this information is

localized or unique to promoters.

#### 1.1.1.1 Interpretation of *in vivo* data

***Basic approach:*** One common way to discover transcription factor binding patterns is through the use of *in vivo* high-throughput data such as chromatin-immunoprecipitation (ChIP-chip). ChIP-chip measures the binding affinity of a transcription factor to regions of DNA. This binding affinity is measured by comparing the amount of DNA bound to a chosen transcription factor in a cell population to the sum total of all the DNA in those cells.

***Current limitations:*** Since this binding affinity averages over a population sample, the relationship of these values and what is going on in an individual cell is unclear. In addition, the resolution of ChIP-chip is about fourty times larger than the actual size of the transcription factor binding patterns. Therefore, one common approach in interrogating ChIP-chip data is to select regions of DNA with an affinity value above a particular threshold and perform statistical evaluation of the DNA patterns within these regions. This approach excludes the information contained in the ChIP-chip binding affinity value.

***Improvements:*** We will present an alternate way to identify DNA sequences associated with the binding of a transcription factor to DNA. This method fully utilizes the binding affinity values produced by a ChIP-chip experiment by integrating them into the DNA sequence analysis.

#### 1.1.1.2 Sequence pattern analysis

***Basic approach:*** In recent years, many computational methods have been developed to detect *de novo* TF sequence motifs. There is a host of freely available motif-finding software which takes DNA sequences as an input and hunts for common elements among those sequences. These methods utilize statistical techniques such as the Gibbs sampler [47], expectation-maximization algorithm [48] and information content [38][79].

***Current limitations:*** As with any statistical analysis, these algorithms are sensitive to the null-hypothesis, or the chosen DNA sequence background, and the the biological significance of the discovered motifs is uncertain without additional wet-lab verification. In addition, a major limitation of many of these algorithms is that there is no standard way to determine whether several similar-looking identified sequence elements represent unique DNA binding sites or if they are a degenerate sampling of the same binding site. Finally, the majority of these algorithms are tuned to find shorter, continuous patterns in the DNA even though genetic control is known to be a complex dynamical process relying on the simultaneous involvement of many transcription factors, suggesting there should also be longer-range structure in the DNA sequences.

***Improvements:*** We develop an algorithm which attempts to address many of the limitations of current *de novo* motif finding algorithms. This algorithm combines concepts from statistics and graph theory to group DNA sequences and is especially tuned to find complex longer-range structures.

### 1.1.2 The Relationship betweeen genes and biological functions

#### 1.1.2.1 Networks: a higher-order interpretation of genetic regulation

Since the late 1990s physicists and applied mathematicians have been impacting the field of biology through the application of network theory to biological systems [85][4]. Network analysis provides a systems-level framework to understand dynamical interactions between genes. These analyses have had a profound effect on our understanding of genetic regulation.

Studying the structure of biological networks has provided new insights into biological functions. The density and degree distribution of regulatory networks have been shown to be indicative of the robustness of the network against aberrant mutations [1][49]. Other structural elements, such as network motifs have been shown to be an important element in many biological networks and may be involved in common, elementary biological functions [74][91]. In addition, many networks are known to have community structure, meaning that there are clusters of nodes in the graph within which there are many edges but between which there are few edges [34]. In regulatory networks such a community may be associated with particular pathways or genetic functions, allowing us to assign biological meaning to the global structure of the network [63].

#### 1.1.2.2 Evaluating and improving gene regulatory networks

***Basic Approach:*** In recent years there has been much excitement about the development of gene regulatory networks and the insight they give into the

organization of genetic pathways. Reconstruction algorithms have been developed which use data such as mRNA expression to evaluate the potential for a regulatory link between pairs of genes [94][55][28]. A handful of algorithms also integrate other various types of information into their reconstruction to further improve the reliability of the produced regulatory network [44][50].

***Current Limitations:*** Despite the development of these reconstruction algorithms, only small number of biological organisms have fully-developed gene regulatory networks. There is often a large amount of noise in the results of reconstruction algorithms, limiting their usefulness in biological applications. Attempts to reconstruct regulatory networks using high-throughput mRNA expression data have been increasingly successful, but the quality of these biological networks still needs improvement. By working under the hypothesis that genes which are involved in many of the same biological functions are likely to be connected in a regulatory network, some reconstruction algorithms have tried to integrate functional information [50][44]. However, there has been little discussion on what the functional similarity of two genes means in a regulatory network context.

***Improvements:*** We explain how to use functional annotation data from the Gene Ontology to give additional biological meaning to links in regulatory networks. This information can also be used to improve network reconstruction.

#### 1.1.2.3 Relationships between biological functions in the Gene Ontology

***Current Approach:*** The Gene Ontology (GO) provides biologists with a controlled terminology with which to describe how genes are associated with function and how those functional terms are related to each other [2]. The terms in the Gene Ontology are organized in the form of a directed acyclic graph (DAG), determined independent of any species information, and has three independent branches: Molecular Function, Biological Process, and Cellular Component. Terms may have multiple parents as well as multiple children but can only belong to one of the three main ontologies. The relationships between terms are normally determined by a collection of individuals in the scientific community and thus reflect a human interpretation of how to classify biological functions, rather than any experimental or computational method [84].

***Current Limitations:*** Because of the DAG structure there are no defined connections between terms assigned to different ontologies, even though there are known biological cases in which a term in one branch is related to a term in another branch. Within an individual ontology, the DAG structure limits the connections between biological functions to those which have a hierarchical relationship. However, it is probable that many terms within the same ontology are related biologically even though they do not have parent/child relationship.

***Improvements:*** We use gene-term annotations found in GO to investigate the relationships between functional terms. Our method allows for the discovery of

functional relationships outside of the established DAG structure.

## 1.2 More Detailed Overview of Projects

In this thesis I will present work done in two very important branches in the field in systems biology: DNA sequence analysis and network theory. Chapters 2 and 3 will describe two distinct approaches to interrogate DNA sequence data. One approach utilizes *in vivo* data in order to give meaning to DNA sequence (Chapter 2) while the other is purely computational and focuses on the higher-order structure of the regulatory regions of DNA (Chapter 3). Chapters 4 and 5 involve an investigation of the graph structure of the Gene Ontology. They explore how the information encoded in this graph can be used to interpret gene regulatory networks (Chapter 4) and to better understand the relationships between biological functions (Chapter 5).

### 1.2.1 Chapter 2: Analyzing ChIP-chip data in the context of DNA sequence

In Chapter 2 we propose a measure that quantifies the binding affinity of specific DNA sequences to a protein utilizing ChIP-chip data. The study focuses largely on data involving RNA Polymerase II (RNAP). RNAP binds to the promoters of housekeeping genes across all cell-types. Housekeeping genes are genes that are typically needed for maintenance of the cell and their promoters are known to often contain CpG Islands, or regions of DNA where the CG dinucleotide occurs with

much greater relative frequency than elsewhere in the genome. However, little detail is known about the specific DNA sequences which compose CpG Islands which may potentiate RNAP binding.

"Nmer-association-with-RNAP" captures the strength of an n-base-pair long DNA sequence's association with promoters bound by RNAP. This measure reveals that virtually all sequences enriched in promoters with high RNAP binding values contain a CpG dinucleotide. Of CpG-containing 8mers, those with the highest association values are primarily variants of six CpG-containing TFBS known to preferentially localize in the proximal promoter. The frequency of these six DNA motifs can predict housekeeping promoters as accurately as the presence of a CpG Island, suggesting that they are the structural elements critical for CpG Island function.

An extension of this analysis which investigates the binding affinity of other proteins in the context of epigenetics is contained in Appendix B.

The work presented in this chapter was done in collaboration with the lab of Dr. Charles Vinson at the National Cancer Institute at the National Institute of Health and was published in *BMC Genomics* in 2008 under the title "All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues." The ChIP-chip experiments were performed by Julian Rosenberg and the data analysis was done in close collaboration with Andrey Shlyakhtenko.

### 1.2.2 Chapter 3: Analyzing DNA sequence for long-range regulatory patterns

In Chapter 3 we develop a *de novo* DNA motif-finding algorithm which centers around the concept of the "split-nmer," by which we mean an $n + k$ base-pair DNA sequence with $k$ base-pairs of degenerate DNA (A, C, G or T) surrounded on both ends by $n/2$ fixed bases. For this study $n = 8$. Split-8mers share many of the statistical frequency properties of 8mers, but allow us to find longer and discontinuous DNA motifs.

The algorithm is illustrated using promoter data from *Drosophila melanogaster*. It is able to correctly identify the majority of canonical promoter DNA binding motifs, including TATA, INR, DPE, DRE, E-Box and others. In addition it identifies many pairs of DNA binding motifs which have preferential spacing. This includes DMv5-DMv4, NDM2-NDM2, INR-DPE, TATA-INR, DRE-DRE and more. In addition, we find many novel promoter elements. These elements tend to localize in the proximal promoter and are found to be involved in developmental biological function, suggesting that these motifs play a cooperative role in the more complex regulation of non-housekeeping genes.

The work presented in this chapter was done in close collaboration with Peter C. FitzGerald of the National Cancer Institute at the National Institute of Health.

### 1.2.3 Chapter 4: Using function to analyze regulatory networks

In Chapter 4 we explore the use of the Gene Ontology in projecting and interpreting gene regulatory networks. Although it is probable that shared function is correlated with the likelihood that two genes are connected in a regulatory fashion, because of the complex structure of the GO DAG, determining how to calculate this functional similarity is non-trivial.

Regulatory information [31] combined with annotation data for *E.coli* indicates that the presence of many shared lowest-level annotations in the Gene Ontology is a good predictor of a regulatory link between two genes. Furthermore, this shared functional annotation is highly correlated with a link's importance to information flow in the established regulatory network. The links predicted using GO annotations are different from those predicted by calculating mutual information from expression. By combining the predictions from the two approaches the accuracy of the reconstructed network can be improved.

The work presented in this chapter was done under the guidance of Michelle Girvan and Edward Ott of the Physics Department at the University of Maryland.

### 1.2.4 Chapter 5: Determining the functional properties of groups of genes

In Chapter 5 we use annotations within the Gene Ontology to investigate whether there exists an alternate logical organization of terms that is different from the DAG. Although there has been some work done on discovering connections

between terms which are not in the DAG [59][77], to our knowledge, there has never been a comprehensive study which investigates whether the structure as a whole is the only legitimate way to classify biological functions.

We used human annotation data to connect functional terms based on shared gene annotations. The communities of functional terms which we found are very different from the branches of the DAG. Cancer signatures are statistically enriched in the found communities of terms, indicating that these communities can provide an alternate natural framework with which to investigate the genetic function of groups of genes. In addition, the classification of terms into communities results in very different partitions in different species, suggesting that these communities may represent a species-specific manner by which to classify biological function.

We also investigated how weighting plays a role in the discovered term communities. This is discussed in Appendix C.

The work presented in this chapter was done under the guidance of Michelle Girvan and Edward Ott of the Physics Department at the University of Maryland.

# Chapter 2

# CpG containing sequences are enriched in promoters bound by RNA polymerase II

Using Chip-chip experimental data from three mouse tissues - liver, heart ventricles, and primary keratinocytes - we determined that 94% of promoters have similar RNAP binding, ranging from well-bound to poorly-bound in all tissues. We combined this data with genomic sequence data in order to evaluate the DNA sequences enriched in the promoters of housekeeping genes. We quantified the association of 8-base-pair long sequences with RNAP binding in multiple tissues through a value we term "nmer-association." A histogram of these values results in a bimodal distribution. Combining this enrichment score with localization information we discovered that variants of six known CpG-containing TFBS can predict housekeeping promoters as accurately as the presence of a CpG Island, suggesting that they are the structural elements critical for CpG island function.

## 2.1 Introduction

The promoter region of genes is typically divided into two regions: the core or basal promoter region and the proximal promoter. The core promoter region stretches from around -50 bp to +20 bp and is the location in the promoter where the pre-initiation complex forms and the general transcriptional machinery assembles,

including RNA polymerase II (RNAP). The proximal promoter extends from -200 bp to the transcriptional start site (TSS) and contains transcription factor binding sites (TFBS) that are critical for the recruitment of RNA polymerase II (RNAP) to DNA [76][57][37]. In mammalian genomes, the CpG dinucleotide occurs at 20% of the expected frequency [81] and is typically methylated both in cell cuture and animal tissues [10][9]. The exception is in CpG islands. CpG islands are defined as regions in the DNA at least 200 bp long where C+G comprise more than 50% of the nucleotides and CpG dinucleotides occur at greater than 60% the expected frequency (this represents roughly 8 or more CpGs in 200 bp) [32]. The presence of CpG islands is associated with gene regulatory regions [43] and in the promoters of genes generally correlates with binding by RNA polymerase II (RNAP) [43]. Promoters of housekeeping genes are constitutively bound by RNAP in all tissues while regulated promoters, either tissue specific or inducible, are selectively bound by RNAP in only certain tissue(s) or contexts respectively [76].

Three advances allow us to interrogate the genome-wide properties of promoters. First is the availability of complete genomic sequences. Second is the determination of full-length cDNAs that can identify the TSS and proximal promoter [16]. Third is the determination of the chromatin architecture of the genome by the identification of hypersensitive sites [72][23] or the location of particular proteins or their modified forms using chromatin immunoprecipitation followed by microarray analysis (ChIP-chip) [69]. Although ChIP-chip experiments have identified the location of RNAP and components of the preinitiation complex in particular tissues [43][5], these experiments have not been done systematically over a range of tissues.

We show that CpG containing DNA sequences are associated with RNAP binding to the same promoter in multiple tissues. Many DNA sequences are more abundant near the TSS than elsewhere [30][8][56][90] and the six most abundant CpG containing sequences that are localized in proximal promoters are known TFBS and can predict RNAP binding to housekeeping promoters with similar accuracy as the presence of CpG islands.

## 2.2 Results

### 2.2.1 Binding of RNAP and H3K9me2 to mouse promoters in keratinocytes, liver, and heart ventricles

To gain insight into the DNA sequence properties of housekeeping promoters, we analyzed RNAP binding to promoters in three mouse tissues: primary skin keratinocytes, liver, and heart ventricles. Using ChIP-chip experiments [89], we determined the genomic localization of initiating (hypo-phosphorylated) RNAP [66][73] in all three tissues (Figure 2.1 A-C). DNA from the RNAP ChIP analysis was amplified and hybridized to Nimblegen mouse promoter microarrays containing 15 probes spanning from -1,000 bp to +500 bp.[1] Signal intensities were averaged for each promoter to produce a number representing binding at each promoter. This produced a graded binding of RNAP to promoter regions as has been previously observed [43][5][35]. We limited the following analysis of DNA sequence properties to the set of 14,790 promoters that contains neither similar/duplicated sequences

[1] ChIP-chip experiments performed by Julian Rosenberg

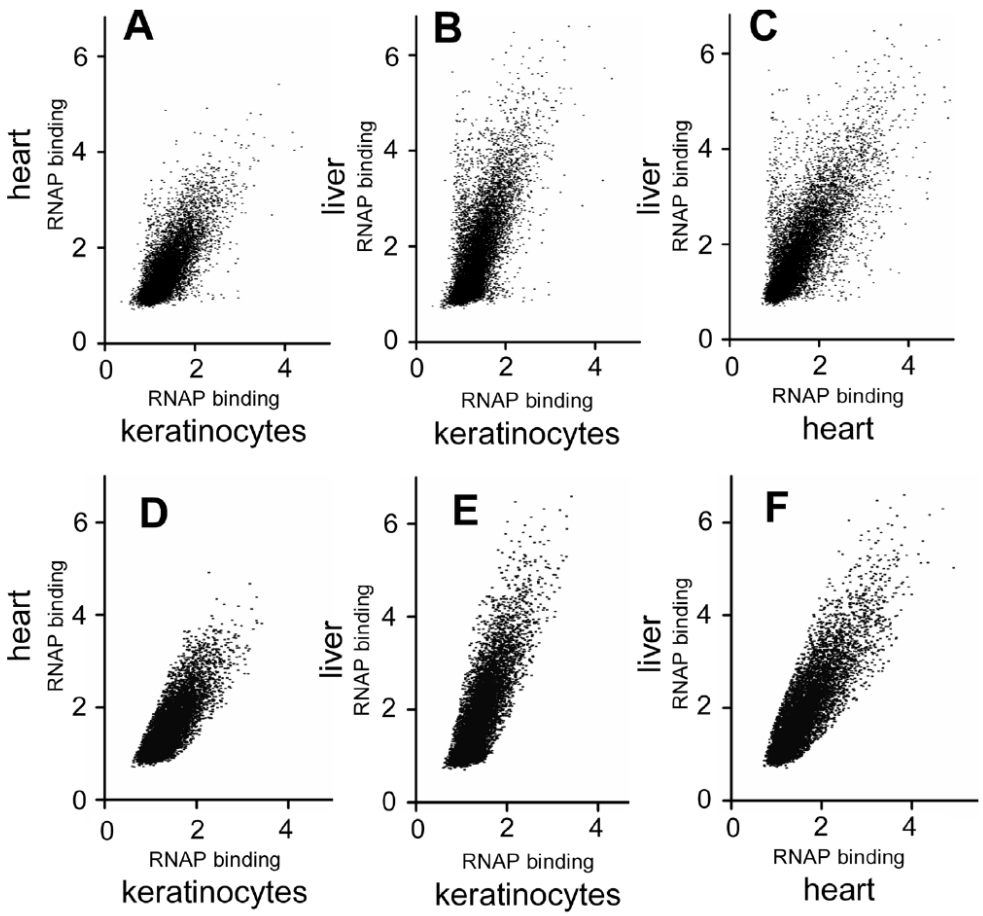



Figure 2.1: A-C) RNAP binding to 14,790 promoters from ChIP-chip data in different mouse tissues with each spot representing a single promoter. A) keratinocytes versus heart ventricles (R = +0.76). B) keratinocytes versus liver (R = +0.73). C) heart ventricle versus liver (R = +0.76). D-F) RNAP binding to the 13,861 promoters with similar RNAP binding values in heart, liver and keratinocytes.

nor a poorly annotated transcriptional start site (TSS).

We then identified promoters that had similar RNAP binding values in all three tissues by excluding genes where RNAP binding between any pair of tissues was significantly different. In order to better compare data sets we performed a data transformation using the two-dimensional rotation matrix. For every pair of experiments, $A$ and $B$ "rotated binding values" were determined by operating on

the original binding values:

$$\begin{vmatrix} b_A^{rotated} \\ b_B^{rotated} \end{vmatrix} = \begin{vmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{vmatrix} \begin{vmatrix} b_A \\ b_B \end{vmatrix},$$

where $\theta$ is the angle by which we rotated the coordinates in the two-dimensional plane and $b$ is the binding value of a protein to a promoter. We rotated each pair of data so that the best-fit line (using least-squares) was the 45-degree line. We forced the best-fit to the origin by subtracting the value of the vertical-intercept of the best-fit line from the vertical data before the rotation. We did this iteratively for every pair of experiments.

In order to assure that the rotation was robust and not heavily influenced by outliers in the data set, we temporarily removed data more than one standard deviation from the original best fit line. If the best-fit line of the transposed data still maintained its 45-degree angle within some small error range, we concluded the data was successfully rotated. If not, then we repeated the procedure using the new rotated values and only those points within one standard deviation of the best-fit line to determine the new rotation angle and intercept adjustment. This was repeated until the best-fit line did not significantly alter with the removal of data points more than one standard deviation from 45 degree line.

Promoters whose rotated binding values were more than two standard deviations off of the 45-degree best-fit line (as determined above) through any of the three pair of data (liver-heart, liver-keratinocytes, and heart-keratinocytes), were consid-

ered "tissue-specific" (not commonly bound). Of our original set of 14,790 promoters, 929 were not commonly bound by RNAP in all three tissues, leaving 13,861 promoters which were commonly bound in all three tissues. 356 promoters which were more than two standard deviations above the best-fit line in liver as compared to heart and keratinotyes were termed "liver-specific-promoters," and similarly, 131 promoters were identified as "heart-specific-promoters," and 47 were identified as "keratinocyte-specific promoters." 395 promoters had high RNAP binding in two of the three tissues.

The remaining 13,861 promoters (94%) have similar RNAP binding in all three tissues, some being well bound by RNAP and others having little RNAP at the promoter (Figure 2.1 D-F). For each of these 13,861 promoters, termed common RNAP promoters, the rotated binding values from the three tissues were averaged, producing a single number representing RNAP binding to a promoter across the three tissues.

To investigate the DNA sequence properties of the 13,861 common promoters (-1,000 bp to +500 bp) and determine potential transcription factor binding sites (TFBS) that are responsible for RNAP binding we analyzed the occurrences of 8 bp-long DNA sequences (8mers) in common RNAP promoters. 8mers were chosen because their length is similar to that of known TFBS. 8mers were counted on the sense and anti-sense strands because, with the exception of TATA [29], 8mers are not restricted to a single strand. Of all 32,896 8mers (38% contain CpG) we extensively characterized the 12,208 most abundant 8mers of which only 20% contained a CpG highlighting that the CpG dinucleotide is underrepresented even in promoter regions

[30].

### 2.2.2 All 8mers enriched in promoters well bound by RNAP in multiple tissues contain a CpG dinucleotide

To measure 8mer enrichment in promoters commonly bound by RNAP, we calculated the term "8mer-association-with-RNAP" for all 8mers. For a particular 8mer, this quantity ($b_8$) is the average RNAP binding to promoters that contain that 8mer, normalized by the average RNAP binding to all common promoters ($\bar{b_p}$).

$$b_8 = \frac{\sum_p b_p \delta_{8p}}{\bar{b_p} \sum_p \delta_{8p}},$$

where $p$ is the promoter in question. $\delta_{8p}$ is equal to one if the 8mer occurs in the promoter sequence and zero otherwise.

A histogram of these values has a bimodal distribution. 20% of 8mers are associated with high RNAP binding to common RNAP promoters (Figure 2.2 A). This result suggests that the graded binding of RNAP to promoters is caused by a combination of 8mers, some of which favor RNAP binding and others which do not. The region of the promoter (-1,000 bp to +500 bp) critical for the observed bimodal distribution extends from -600 bp to +400 bp (Figure 2.2). Strikingly, nearly all the 8mers that are associated with RNAP binding contain the CpG dinucleotide while virtually none of the remaining 8mers contain a CpG. In contrast to the CpG dinucleotide, the other dinucleotides did not exclusively occur in either part of the

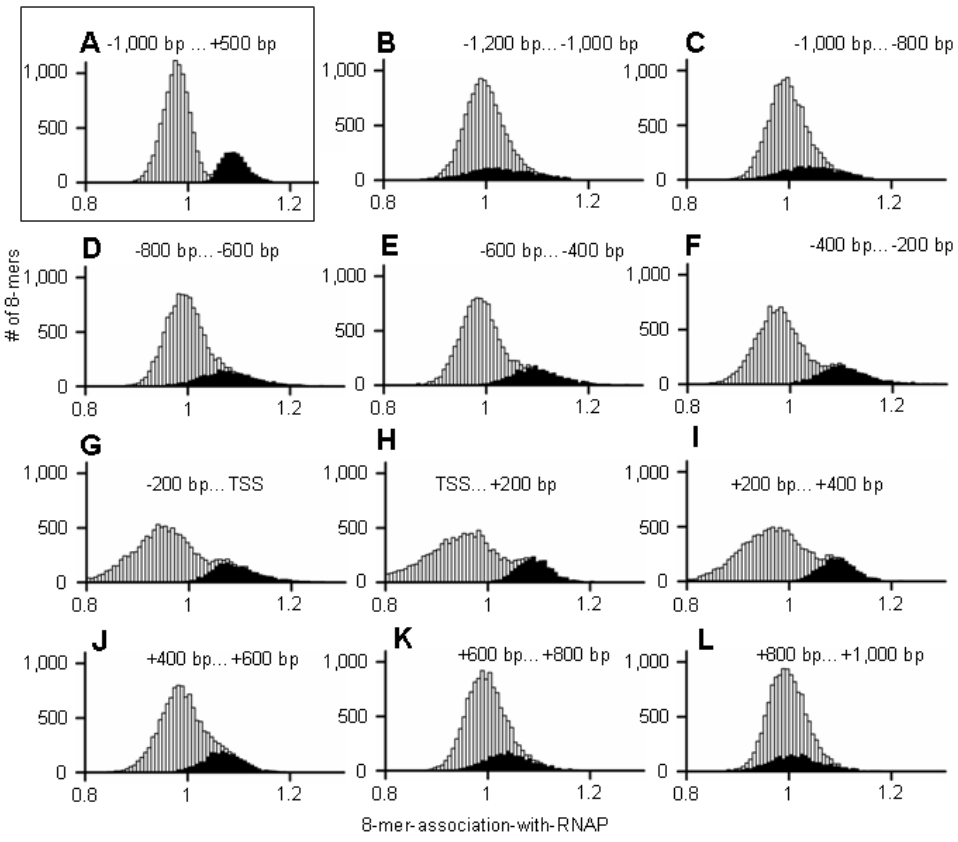



Figure 2.2: A) Histogram of the 8mer-association-with-RNAP between -1,000 bp and +500 bp for abundant 8mers in the 13,861 common RNAP promoters. 8mers that contain a CpG are colored in black. B-L) Histogram of the 8mer-association-with-RNAP in 200 bp increments from -1,200 bp to +1,000 bp. 8mers that contain a CpG are colored in black.

bimodal distribution (Figure 2.3).

To evaluate if other types of promoters have a different enrichment of 8mers, we examined the transcriptionally inactive genes marked by a post-translationally modified form of histone 3, H3K9me2 (lysine 9 containing a dimethyl group) [62][51]. In keratinocytes, ChIP-chip identification of H3K9me2 genomic localization negatively correlated with RNAP (correlation coefficient, R = -0.50) (Figure 2.4 A). The 8mer-association-with-H3K9me2 also had a bimodal distribution with the CpG

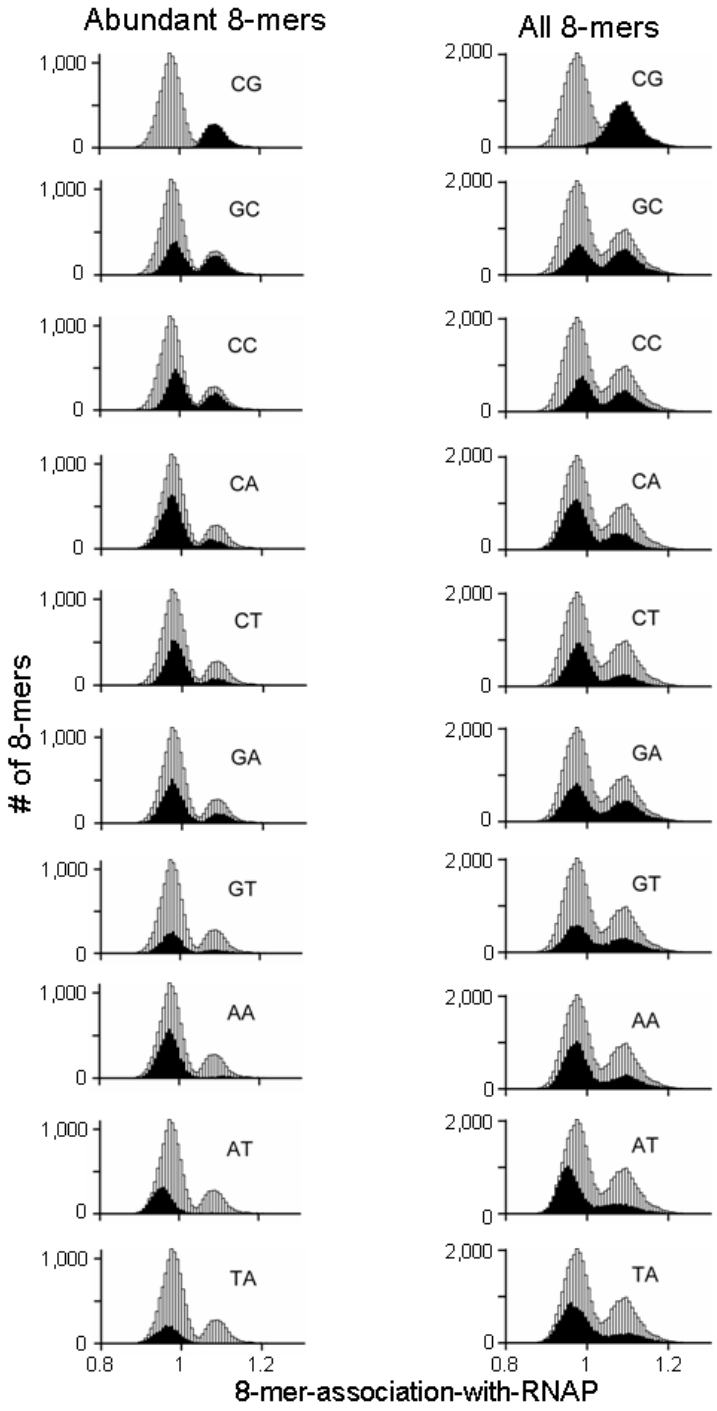



Figure 2.3: Histogram of the 8mer-association-with-RNAP between -1,000 bp and +500 bp for abundant 8mers with 8mers containing each of the 10 dinucleotides noted in black.

containing 8mers associating the least with H3K9me2 binding (Figure 3B). As anticipated (comparing Figure 2.2 A and Figure 2.4 B), practically all the 8mers most associated with common RNAP binding also are least associated with H3K9me2 binding (Figure 2.4 C).

The 8mers without a CpG were also plotted separately to highlight the few 8mers that are the exception to the general conclusion that only CpG containing sequences are associated with RNAP binding to a promoters (Figure 2.4 D). The most notable exception is the GACCAATC 8mer, a CCAAT sequence that is enriched in housekeeping promoters.

Previous work indicated that ∼50% of human promoters bound by RNAP contain the INR and DPE consensus sequences between -200 bp and +200 bp [43]. To see if these non-CpG-containing sequences were also exceptions to our general conclusion, we calculated the association-with-RNAP and association-with-H3K9me2 for TATA, INR and DPE in the set of promoters with similar RNAP binding values in the three tissues we have examined. This was accomplished by averaging the binding values of those promoters that contained the consensus sequence at the expected position [57]. In mouse, the consensus TATA is uniquely positioned in only 1.8% of promoters and has a very high association-with-H3K9me2 binding to promoters. The INR was uniquely positioned in only 9% of promoters and is associated with H3K9me2 bound promoters. DPE is not uniquely positioned in promoters, but occurs in 19% of promoters at the expected location and is also associated with H3K9me2 binding (Figure 2.4 C). This suggests that TATA, INR and the DPE are not important for RNAP binding to promoters in multiple tissues. Presumably

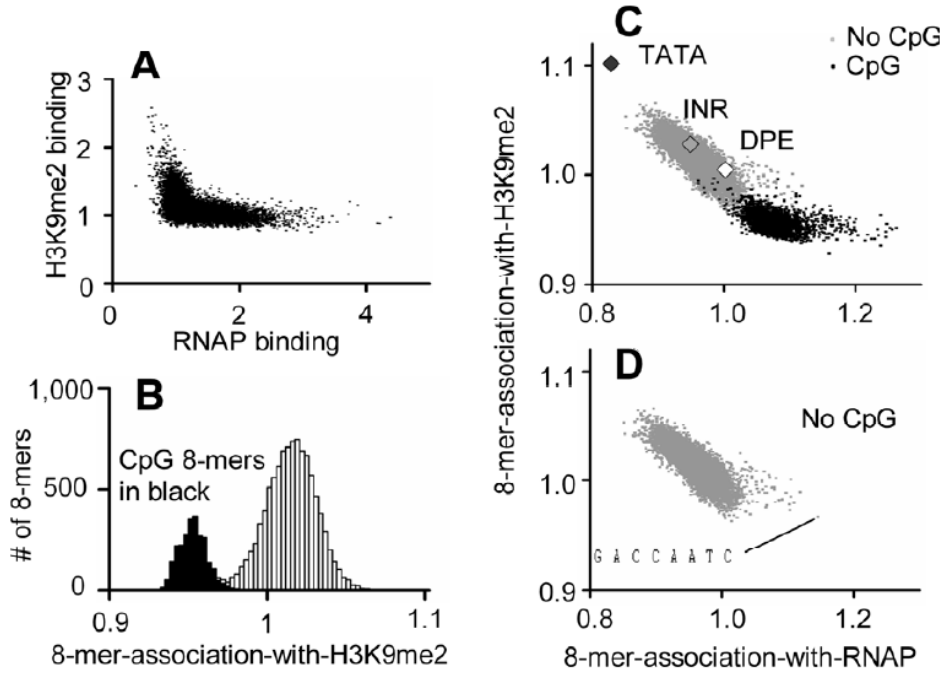



Figure 2.4: A) Binding of RNAP vs. H3K9me2 (R = -0.50) in mouse tissue culture keratinocytes. B) 8mer-association-with-H3K9me2 for 12,208 abundant 8mers, calculated for 14,790 promoters between -1,000 bp and +500 bp; CpG containing 8mers are colored in black. C-D) 8mer-association-with-RNAP vs. 8mer-association-with-H3K9me2. C) All 8mers. The association-with-RNAP and the association-with-H3K9me2 for the core promoter elements at their unique position in promoters is presented for TATA (TATAWAAR), INR (YYANWYY) and DPE (RGWYV). D) 8mers without a CpG.

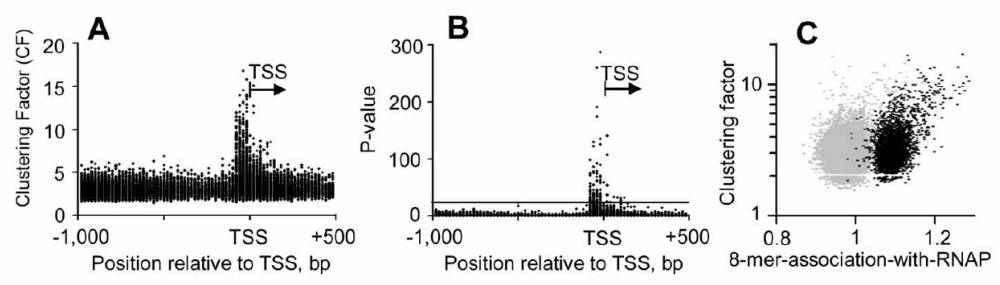



Figure 2.5: A) A measure of non-random distribution termed a Clustering Factor (CF) is plotted in the most populated bin for 8mers with at least 20 members in the most populated 20 bp bin (abundant 8mers). Note the dots between -100 bp and the TSS with large CF values representing 8mers that are more abundant near the TSS than elsewhere. B) A probability term P for the 8mers in (A). A P-value of 24 means that the distribution of the 8mer has a less than $10^{-24}$ chance of being random. C) Non-random distribution of 8mers (Clustering Factor) vs. 8mer-association-with-RNAP for abundant 8mers.

these sequences are important for tissue-specific gene expression.

### 2.2.3 Non-random distribution of 8mers in promoters

If the 8mers that associate with RNAP binding are TFBS, they may be localized in the proximal promoter as has been observed in human [30][8] and Drosophila promoters [29]. We thus determined the "Clustering Factor" (CF, a measure of non-random distribution between -1,000 bp and +500 bp) [30][29] for abundant 8mers in promoters, and compared it to 8mer-association-with-RNAP. Some 8mers were preferentially localized near the TSS (Figure 2.5 A-B). The 8mers most associated with promoters commonly bound by RNAP had a high CF (Figure 2.5 C). However, there was also a class of 8mers with high CFs but low 8mer-association-with-RNAP values that may represent TFBS involved in regulated gene expression.

The 120 8mers with the statistically highest CF (Figure 2.5 B) that localize

| Motif | Sequence | 8mer-association-with-RNAP |
|---|---|---|
| BoxA | TCTCGCGA | 1.30 |
| NRF-1 | GCGVTGCG | 1.24 |
| ETS | VCCGGAARY | 1.21 |
| CRE | TGACGTCA | 1.19 |
| SP-1 | CCCCGCCC | 1.14 |
| E-Box | YCACGTGA | 1.10 |
| CCAAT | RRCCAATSR | 1.04 |
| KLF | CCCCTCCC | 1.04 |
| TATA | TATAAAD | 0.96 |
| CRE-T | TGATGTCA | 0.90 |

Table 2.1: Association of the ten localized motifs with RNAP binding.

upstream of the TSS could be manually grouped into ten consensus motifs that are known TFBS: ETS, NRF-1, E-Box, BoxA, CRE, SP1, KLF, CCAAT, TATA, and CRE-T, six of which contain a CpG dinucleotide (ETS, NRF-1, E-Box, BoxA, CRE, and SP1). A similar analysis has identified that these ten motifs localize to the proximal promoter in human promoters [30]. To see if these TFBS play some specific role in RNAP binding, we calculated the association-with-RNAP for the consensus sequences of these TFBS (Table 2.1). As expected, the CpG-containing TFBS have high association values for RNAP binding. ETS, NRF-1, and BoxA correlate the best with RNAP binding to promoters in multiple tissues.

### 2.2.4 CpG Islands can be defined by two or more of the six CpG containing TFBS.

Previous work has suggested that housekeeping genes can be defined by the presence of a CpG Island in the promoter region [32], but the DNA sequences

properties of CpG Islands has not been described. We evaluated if the presence of the six CpG consensus motifs in proximal promoters (-200 bp to the TSS) predicts RNAP binding to promoters commonly bound by RNAP and compared these results with the occurrence of a CpG Island between -200 bp to the TSS (Figure 2.6 A). The results demonstrate that the presence of any two of these motifs recapitulates the discrimination based on the presence of a CpG Island in regards to RNAP binding to common promoters. In order to compare these two measures, we grouped promoters into ten equal size groups with increased RNAP binding. 80% of promoters in the group best bound by RNAP contain a CpG Island and a similar number contain two or more of the six motifs (Figure 2.6 A). Similarly, only 5% of promoters with the lowest RNAP binding values are CpG Islands, and only about 5% have two or more motifs (Figure 2.6 A). The presence of three or more of these motifs produced a lower positive hit rate in the best bound group (48%) but occurred in only 1% of promoters not bound by RNAP. Therefore, our analysis suggests that CpG Islands have predictive value in defining housekeeping genes because of the presence of these six TFBS motifs. These six motifs do not account for all CpGs in CpG Islands. Some of the other CpGs are known TFBS but the function of the rest remains unclear. They could be sequences that persist because they are protected from methylation and ultimate destruction or they could be involved in the higher-level regulatory processes that have been proposed for CpG Islands [42]. In contrast to promoters well bound by RNAP in multiple tissues, only 20% of tissue specific proximal promoters are CpG Islands and similarly only 20% contain two or more of these six motifs. This indicates that these six motifs correlate with promoters that

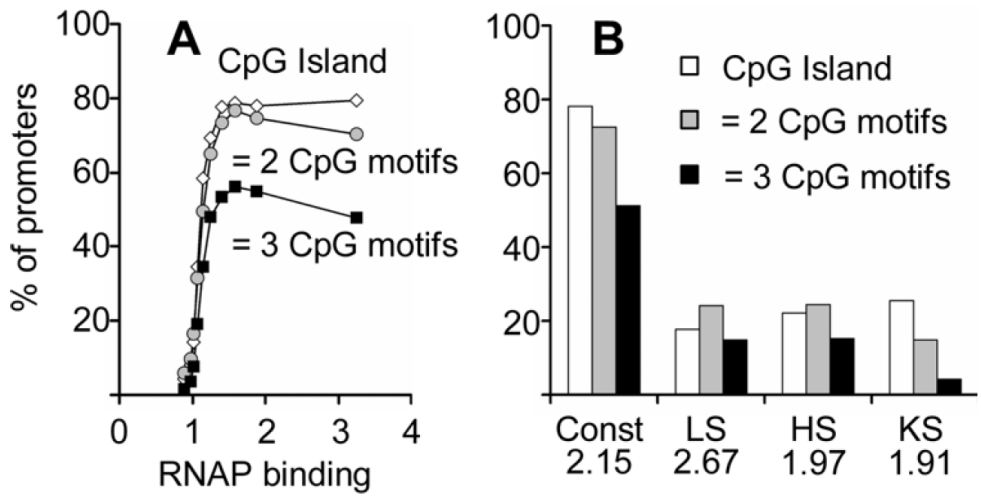



Figure 2.6: A) Fraction of promoters that contain particular sequences between -200 bp and TSS: 1) CpG Island, 2) two or more of six CpG containing motifs (SP1: CCCGCC, CCGCCC, CGCCCC; ETS: CCGGAA, GCGGAA; NRF-1:CGCATGCG, CGCGTGCG, CGCCTGCG; BoxA: TCTCGCG, CTCGCGA; CRE: ACGTCA; E-Box: CACGTG), 3) three or more of the six motifs. B) Fraction of promoters that contain particular motifs: top 20% of common RNAP promoters (Const), liver specific (LS), heart ventricle specific (HS), and keratinocyte specific (KS) promoters. Average RNAP binding for each class is presented.

are bound by RNAP in multiple tissues and not tissue specific promoters (Figure 2.6 B).

## 2.3 Conclusions

We identified promoters that are bound similarly by RNAP in multiple tissues and determined the association between the presence of 8mers in these promoters and the extent of RNAP binding to the promoter. Looking at RNAP binding to housekeeping promoters, we observed a bimodal distribution: only 8mers with the CpG dinucleotide are in the class of sequences most associated with RNAP binding and only 8mers without a CpG are in the class least associated with RNAP binding.

An implication of this observation is that knowing if a TFBS contains a CpG reveals aspects of its biological function. If the TFBS contains a CpG, it is involved in constitutive gene expression and if the TFBS does not contain a CpG, it is involved in regulated gene expression. This insight will help identify potential functions for transcription factors when their TFBS is identified. Additionally, if a transcription factor shows degeneracy in its TFBS [7][67], binding to a CpG sequence and a similar sequence without a CpG, it suggests that this transcription factor is involved in both constitutive and regulated gene expression. This is observed for the CRE and CRE-T sequences, two sequences that are localized in the proximal promoter and differ by a single base: CRE contains a CpG (TGACGTCA) while CRE-T does not (TGATGTCA). The CREB protein binds both sequences well (data not shown) but the two sequences correlate very differently with RNAP binding suggesting that the CREB transcription factor can regulate either constitutive gene expression by binding the CRE sequence or regulated gene expression by binding the CRE-T sequence.

In vertebrates CpG dinucleotides are rare and usually are methylated on the cytosine but do occur at close to the expected frequency in clusters called CpG Islands where the CpGs remain unmethylated [42][11]. These CpG Islands often occur in promoters of housekeeping genes [32][43]. We show that the presence of two or more of any of the six CpG containing TFBS (SP1, ETS, NRF-1, CRE, E-Box, and BoxA) in the proximal promoter can predict RNAP binding to housekeeping promoters as accurately as the presence of a CpG Island in the proximal promoter. Methylation of the CpG in the TFBS has been found to inhibit the DNA binding

for five of the six TFBS that are abundant and localize in proximal promoters suggesting this may be a general result for CpG containing TFBS. Methylation dependent inhibition of transcription factor binding to DNA has two implications. First, the transcription factors that are critical for the activation of housekeeping genes solve the problem of finding their TFBS in the genome by only binding to unmethylated TFBS. Since most CpGs in the genome are methylated, the only places these transcription factors can bind are in the unmethylated CpG Islands in promoters. Second, the pathological methylation of CpG dinucleotides in CpG Islands, as occurs in many cancers [42], would prevent these abundant transcription factors from binding their TFBS thus causing the promoters to become inactive. This could be a critical initial step that subsequently allows CpG methyl binding proteins to bind to methylated CpGs and actively repress a promoter [12].

## 2.4 Further Discussion

In this analysis we primary investigated two proteins, RNAP and H3K9me2, which have broad-based sequence affinities. However, most transcription factors, including some of the ones mentioned as part of the typical RNAP complex, have much more specific DNA-binding preferences. These DNA-binding preferences could quickly be computationally assessed with the 8mer-association calculation. As opposed to the bimodals observed for RNAP and H3K9me2, in the case where a TF has a very specific sequence preference, we would expect a fairly Guassian distribution of 8mer-association with the sequences to which the TF binds as outliers

relative to this distribution.

Although in CpG Islands and the promoters of house-keeping genes is it believed that the CpG dinucleotide is typically unmethylated, in other promoters and other regions of the genome CpGs are believed to be mostly methylated and gene expression regulated in part by epigenetic mechanisms. One shortcoming of the 8mer-association calculation is an implicit assumption that a unique DNA sequence should function similarly independent of its location in the promoter or genome. However, if in one promoter a DNA sequence contains a methylated CpG and in another that DNA sequence is unmethylated, this one sequence may behave very differently. Furthermore, it is believed that the function of DNA sequence may vary with its position relative to the TSS.

For a follow-up analysis which addresses some of the epigenetic properties in DNA sequence analysis for several different transcription factors, see Appendix B.

## 2.5 Acknowledgements

The work presented in this chapter was done in collaboration with the lab of Dr. Charles Vinson at the National Cancer Institute at the National Institute of Health and was published in *BMC Genomics* in 2008 under the title "All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues." The ChIP-chip experiments were performed by Julian Rosenberg and the data analysis was done in close collaboration with Andrey Shlyakhtenko.

# Chapter 3

# A novel motif-discovery method for finding co-occurring or discontinuous DNA motifs

Although many *de novo* DNA motif-finding algorithms do a good job of determining individual transcription factor binding sites (TFBS), they generally are not suited for the discovery of longer range structure that result from multiple DNA interactions with the transcriptional complex. Here we propose a novel motif-finding algorithm which systematically determines all enriched pairs of DNA sequence motifs and begins to address the issues of longer range DNA structure and complex regulation. By using "split" DNA motifs, our algorithm can not only find known regulatory motifs, but in a very simple manner, discover if any sets of motifs preferentially localize relative to one another.

The algorithm is illustrated using promoter data from *Drosophila melanogaster*. We are not only able to correctly identify the majority of known binding motifs, including TATA, INR, DPE, DRE and E-Box, but we also find many pairs of DNA motifs which have preferential spacing. Some examples include DMv5-DMv4, NDM2-NDM2, INR-DPE, TATA-INR, and DRE-DRE. Although several of these pairs have previously been reported to preferentially co-occur, the existence and/or amount of any preferential spacing between these pairs has not been reported. In addition to known elements, we also identified several novel promoters elements,

some which occur individually and others which preferentially localize relative to a known TFBS. These findings suggest that the split-motif algorithm is a powerful tool useful for finding novel promoter elements and for identifying long-range structure in promoters.

## 3.1 Background and Motivation

Gene regulation is controlled, at least in part, by a region of DNA sequence called the core promoter, located upstream of the transcriptional start site (TSS) of each gene. The regulatory involvement of distal regulatory regions of DNA, enhancers, is also increasingly being recognized but their exact properties are still incompletely understood. The core promoter typically contains the majority of DNA binding sites necessary for regulatory proteins such as transcription factors (TF) to bind to the DNA and initiate (or prevent) transcription of a gene. The functionality of these DNA binding sites has been shown to correlate with their position relative to the TSS [83] and is also likely to correlate with their position relative to other binding sites [14].

In recent years, many computational methods have been developed to detect *de novo* TF sequence motifs [41]. Representatives of these methods are CONSENSUS [79][38], EM-based algorithms [48], Gibbs sampler [47], "AlignACE" [70], and "BioProspector" [52]. In addition, there is a host of freely available motif-finding software which takes DNA sequences as an input and hunts for common elements among those sequences. Some of the most popular ones include MEME [3], Al-

legro [36], Clover [75], Weeder [64] and FIRE [27]. Some of these programs can also incorporate *in vivo* data, such as ChIP-chip and microarray values, into their motif-finding algorithms.

The majority of these algorithms are tuned to find discrete transcription factor binding sites (TFBS) which span 5-20 base-pairs (bp). However, there is a substantial body of evidence emerging which indicates that sets of DNA motifs play a biological role in transcription factor binding and gene expression [14][68]. Even so, the issue of complex regulation, the case in which multiple DNA sequence motifs act in concert to regulate gene expression, has yet to be systematically studied. To address this issue we developed an algorithm which systematically searches for all pairs of DNA sequence motifs as well as provides a novel method by which to search for single *de novo* TF binding motifs.

One potential weakness of novel motif finding algorithms is the inability to correctly determine whether several similar-looking identified sequence elements represent unique DNA binding sites or if they are a degenerate sampling of the same binding site. In order to overcome this we centered our algorithm around the co-occurence of DNA sequences. We worked under the hypothesis that two sequence elements which occur together many times in a given sequence set represent two samples of the same DNA binding site, whereas two sequence elements which do not often co-occur are representative of unique binding sites, even if they have a high sequence similarity. Our algorithm, utilizing the concept of a split-motif is demonstrated using a set of 9,494 *Drosophila* promoters.

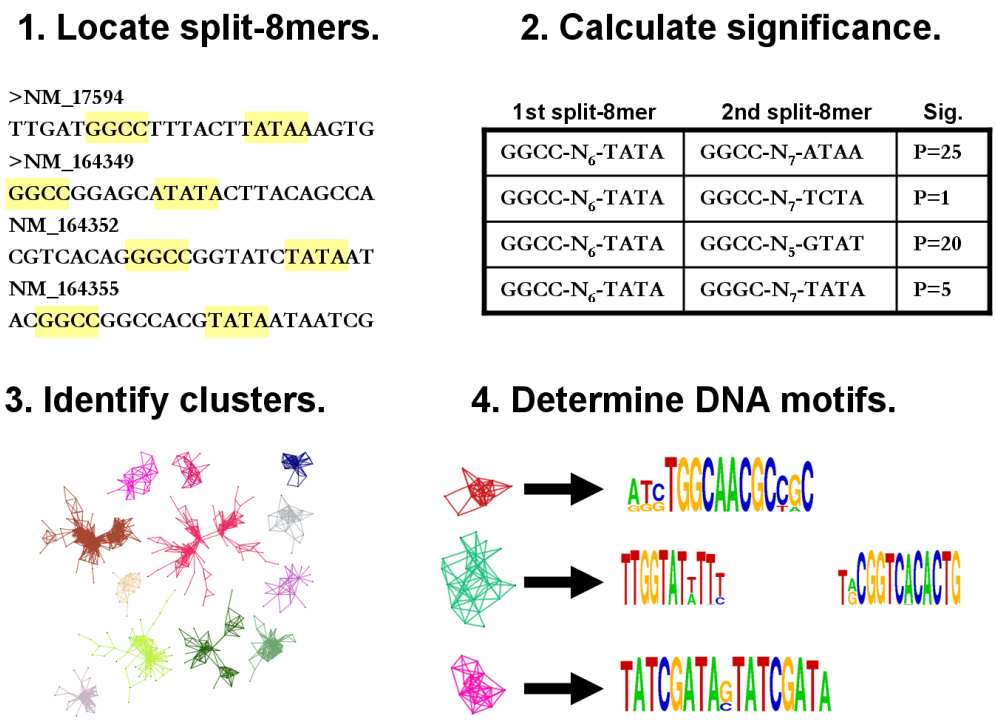



Figure 3.1: Overview of the Split-nmer Algorithm

## 3.2 Methods: Determining the DNA motifs over-represented in an input set of sequences using a split-motif algorithm

The split-motif algorithm (Figure 3.1) systematically calculates the statistical probability of the co-occurrence for pairs of 4mer-$N_k$-4mer, or split-8mers, given an input set of DNA sequence. Pairs of split-8mers whose co-occurrence is statistically higher than expectation are then linked together in a correlation network and consensus motifs for the resulting groups are determined. Biological function can be ascribed to these motifs by identifying which genes in the input data set contributed most strongly to the motifs and analyzing those genes' biological role in the cell.

### 3.2.1 Split-nmers

Our method centers around the concept of a "split-nmer" by which we mean a $k+n$ letter word, where the $n/2$ base-pairs on either end of the word are held fixed and $k$ base-pairs in the center of the word are allowed to be completely degenerate. In this paper we will focus on split-8mers, containing a variable number of degenerate spacer base-pairs surrounded by four unique and fixed bases on either end of the sequence. We chose to focus on split-8mers because the four bases on either end of the sequence should contain enough information to capture the core of many TFBS. Also, the occurrence of the split-8mers should be similar to that of continuous 8mers and be high enough to have statistical viability.

One way to represent a split-8mer is in the form $XN_kY$, where $X$ is the first four-letter word, $Y$ is the second four-letter word, and $N_k$ represents $k$ base-pairs of completely degenerate DNA (A, C, G or T). It is interesting to note that for some pairs of $X$ and $Y$, $k$ can take on a negative value as low as $-n/2$, or in the case of split-8mers, $-4$, representing the occurrence of sequences which are less than $n$ base-pairs in length. For example, the split-8mer TGAC $N_{-2}$ ACGT would actually represent the six-letter word TGACGT. $k$ can therefore vary from $k_{min} = -n/2$ to $k_{max}$, the maximum value of $k$ investigated, representing all split-nmer words with a mimimum length of $n/2$ and a maximum length of $k_{max}+n$, or from 4 to $k_{max}+8$ in the case of split-8mers. These words are either completely unique, as in the case of $k \leq 0$, or have exactly $n = 8$ letters fixed. In our following analysis we set $k_{max} = 55$ as this limit captures the majority of information in our input promoter set.

### 3.2.2 Determining statistically significant pairs of split-8mers

We determined the statistical probability of the co-occurrence of every pair of split-8mers using the hypergeometric probability distribution. This was done by counting the number of sequences in which the first split-8mer occurred ($N_1$), the number of sequences in which the second split-8mer occurred ($N_2$) as well as the number of sequences in which both halves of the two split-8mers occur within a $K$ bp window of one another ($N_{12}$). We imposed the window constraint to better assure that we only counted co-occurrences of two split-8mers which are likely to be biologically relevant because of their physical proximity to one another.

After determining the co-occurrence count $N_{12}$ we modified the value by subtracting $E(N_{12})$, a measure of how many co-occurrences of the two split-8mers one might expect by chance given their sequence similarity.

$$N'_{12} = N_{12} - E(N_{12}), \qquad E(N_{12}) = max[N_1, N_2] \times (0.25)^{n'-b},$$

where $N'_{12}$ is the modified co-occurrence, $n'$ is the minimum number of letters held fixed in the two split-nmers ($n' = n = 8$ for split-8mers unless $k < 0$) and $b$ is the number of shared base-pairs. This modification to $N_{12}$ is crucial to prevent the counting of co-occurrences due only to sequence similarity. As an example, the two split-8mers TGAC $N_{10}$ TGAC and TGAC $N_{11}$ GACT could be aligned to share seven base-pairs in common. This alignment would result in the sequence TGAC $N_{10}$ TGACT, which is only one base-pair different from either of the two original split-8mers. The expected occurrence of the aligned sequence is therefore

equal the occurrence of one of the split-8mers multiplied by the probability that that split-8mer will be expanded by the additional base-pair.

For simplicity, we took the split-nmer with the maximum occurrence to be the one expanded. We reasoned that taking the maximal value of $N_1$ and $N_2$ would "punish" $N_{12}$ more and thus, even if we lost some true positives, would be more robust against producing false positives. For computational ease we also assumed the probability of extending any sequence by one nucleotide is equal to 25%. Since 25% is an approximation, we will use our final p-values for ordering by significance rather than a well-defined probability.

Once we determine $N'_{12}$, $N_1$, $N_2$ and the total number of input sequences ($T$), we can use the hypergeometric probabilty distribution to determine the probability of obtaining $N'_{12}$ or more co-occurrences of the two split-8mers by chance:

$$p = \sum_{N_v=N'_{12}}^{min[N_1,N_2]} \frac{\binom{N_1}{N_v}\binom{T-N_1}{N_2-N_v}}{\binom{T}{N_2}}.$$

We will be taking $P = -log_{10}(p)$ as the p-value in the following discussion.

### 3.2.3 Determining significant DNA motifs represented in the input data set

After calculating the p-value, ($P$), for every pair of split-8mers we proceeded to group together those split-8mers whose pair-wise co-occurrence is statistically significant. This was done by taking all pairs of split-8mers whose p-value is above a particular threshold, using those pairs to construct a correlation network and then identifying communities, or groups of sequences, in the correlation network. We chose our p-value cutoff in such as way as to maximize the number of informative modules in the correlation network. We did this by choosing the p-value cutoff which minimizes the percentage of identified split-8mers found in the largest connected component. This optimization forces the information in the correlation network to be spread amoung many smaller modules. Once the groups of DNA sequences are determined, the sequences within each group can be aligned to produce DNA sequence motifs representing the collection of split-8mers found within the group.

At this point it is beneficial to consider how an idealized split-motif would be represented in the correlation network. As an example, take a perfect 6mer DNA motif which is preferentially localized with respect to another 6mer DNA motif and whose sequence make-up is completely unique, meaning that no part of either 6mer is preferentially localized relative to some other DNA sequence. In this case we would expect all split-8mers represented within the split-motif to statistically co-occur and be connected to each other in our correlation network, and to not be connected to any other split-8mer in the correlation network. This is visualized in

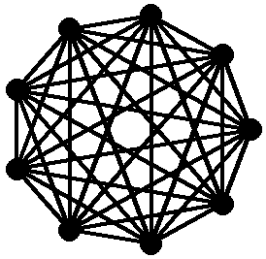


Figure 3.2: Idealized 6mer-$N_k$-6mer as it would look in a correlation network

Figure 3.2. Three 4mers are contained within the first 6mer and three 4mers are contained in the second 6mer, leading to three times three or nine total split-8mers represented in this 6mer-$N_k$-6mer. Some various nmer-$N_k$-mner combinations are shown in Table 3.2.3.

### 3.2.4 Ascribing biological function to the identified DNA motifs

| m \ n | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|
| 4 | 1 | 2 | 3 | 4 | 5 |
| 5 | | 4 | 6 | 8 | 10 |
| 6 | | | 9 | 12 | 15 |
| 7 | | | | 16 | 20 |
| 8 | | | | | 25 |

Table 3.1: Number of split-8mers in a cluster representing an nmer-$N_k$-mner split-motif. These values are equal to $(n-3)(m-3)$.

Once motifs enriched in a set of sequences are identified it is important to revisit the input sequence set and determine the origin of the final motifs within this set. To address these issues we determine which of the input sequences most stronger contributed to forming each cluster in our correlation network. Each edge in the correlation network represents a subset of the input sequences, namely, those input sequences in which the two split-8mers forming the edge both occur. Therefore, by determing these subsets for all edges within a cluster, we can determine how many times each input sequence contributed to the formation of that cluster. Once the input sequences which contributed to a split-motif community are determined, the biological properties associated with those input sequences can be assessed.

## 3.3 Results: split-motifs in *Drosophila* promoters

Using 9,494 *Drosophila* promoters, we calculated the p-value ($P$) for every pair of split-8mers which had a modified co-occurrence of at least 10. We then identified the connected components with at least nine members for various p-value cutoffs and determined the optimal cutoff (see Section 3.2.3). For our input set, this cutoff was 28, however, any cutoff above 15 would give similar results.

At this cutoff we identified seventy-two connected components (Figure 3.3). The members of each component were aligned and represented by a single DNA motif. Of these seventy-two identified motifs, approximately half are shorter "continuous" DNA motifs, and the other half are longer "discontinuous" motifs. The majority of the continuous DNA motifs are known TFBS and the discontinuous motifs are primarily composed of two known TFBS located at different spacings relative to one another. There is also a handful of identified longer continuous motifs which are composed of two shorter known TFBS that are separated by 0bp or are slightly over-lapping. Five continious and discontinuous motifs as well as two longer continuous motifs from the largest components are shown in Table 3.2.

### 3.3.1 Continuous Motifs

Many of the identified continuous motifs are known promoter elements for *Drosophila* and include TATA, INR, DPE, DRE, E-Box and others. However, we also identified six novel elements. We determined which of our input sequences contained 6 or more split-8mer pairs from each connected component and defined these

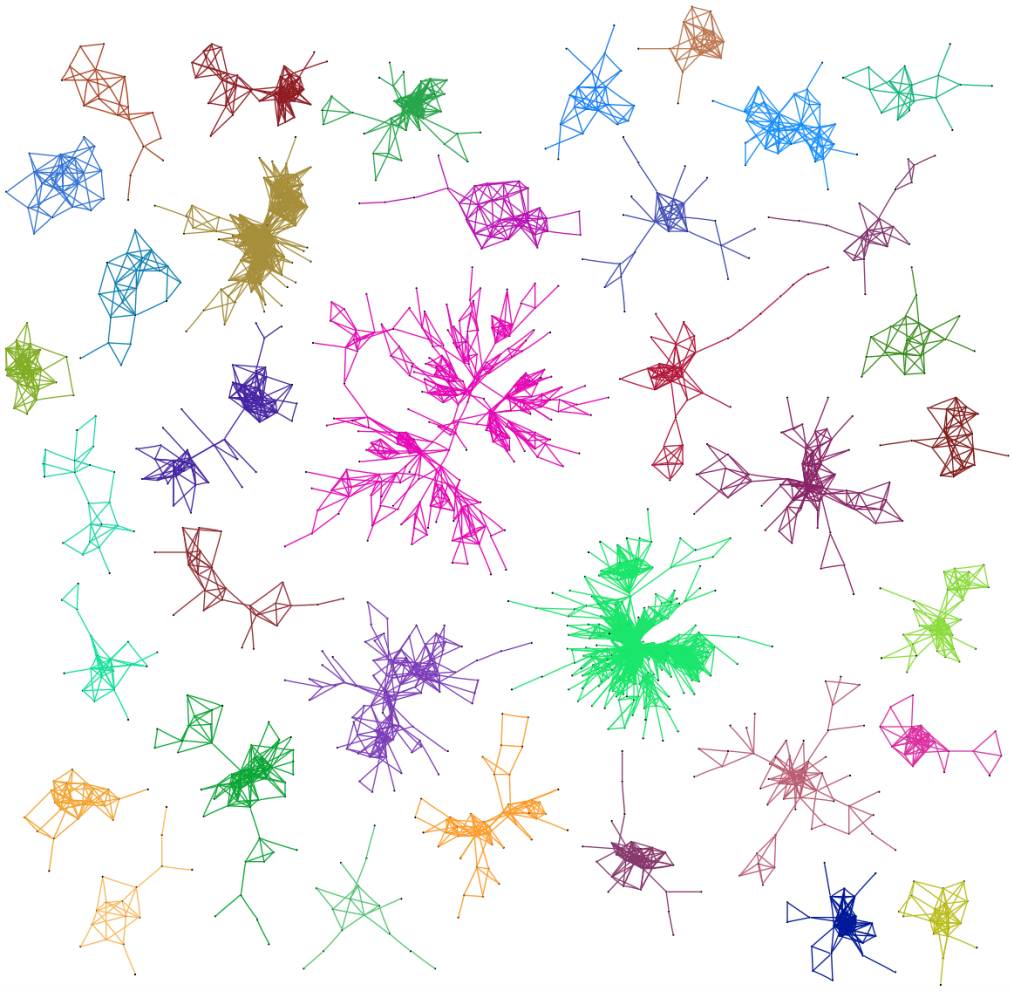


Figure 3.3: A visual representation of the correlation network for *Drosophila* (statistical cut-off value of 28). Each point represents a split-8mer and lines represent a pair of split-8mers whose co-occurrence is statistically significant. Notice how well the diagram falls into idividual connected-components. Each connected component represents a set of split-8mers which can be combined to create a particular DNA motif significant for *Drosophila*.

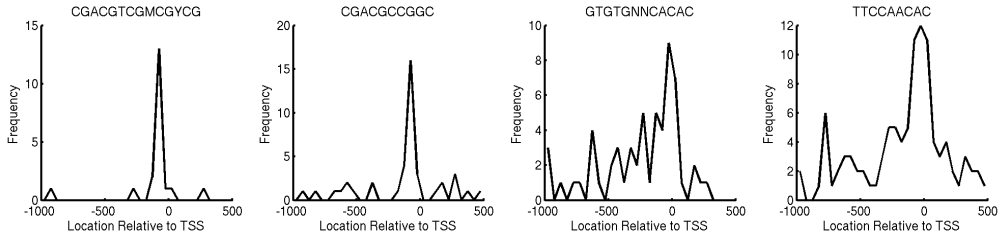



Figure 3.4: The frequency distribution of four of the novel elements discovered by the split-motif algorithm. The fact that these elements peak in the proximal promoter lends support to the idea that these elements are real, biologically relevant components of *Drosophila* promoters.

input sequences as influential in the formation of that component. 207 of our input promoters are associated with one or more of the identified novel elements and none of the other identified known regulatory DNA motifs, indicating that there is a set of promoters which may be regulated by these DNA motifs. These genes are over-represented in developmental GO categories such as "appendage development", "nervous system development", "cytoskeletal protein binding" and "tubulin binding."

We also investigated the localization properties of these six novel elements and found that four of them peaked in the proximal promoter (Figure 3.4). Motif localization in the proximal promoter has previously been shown to be indicative of DNA sequence function [30][29], lending further support that these discovered motifs are biologically important. None of the motifs are especially abundant, which may be one reason why they have not been discovered by other traditional motif-finding algorithms.

Despite the fact that the split-motif algorithm is specifically tuned to find longer and split DNA sequences, it does surprisingly well at finding shorter contin-

| Name | Motif | Component Size | Corresponding Genes |
|---|---|---|---|
| DMv4 |  | 256 | 2289 |
| DRE |  | 255 | 3098 |
| E-Box |  | 90 | 1019 |
| DMv2 |  | 61 | 381 |
| DMv1 |  | 50 | 430 |
| DMv5-DMv4 |  | 104 | 684 |
| INR-DPE |  | 45 | 445 |
| TATA-INR |  | 33 | 295 |
| INR-DPE1 |  | 30 | 253 |
| DRE-DRE |  | 30 | 235 |
| DRE-DRE |  | 51 | 293 |
| NDM2-NDM2 |  | 51 | 201 |

Table 3.2: Examples of motifs identified by the split-motif algorithm. Information is shown for some of the largest identified components.

uous motifs. There were very few known *Drosophila* TFBS which we did not find. One TFBS which did not surface in our analysis is the 7bp-long NDM3. One potential reason this TFBS was missed was that it shares sequence properties with the TATA motif. Since split-8mers only appear once on our correlation network, they will not ever be grouped with two distinct TFBS. The core of NDM3, AAAG, has similar properties to TATA. However, TATA is the more prominent motif, so the split-8mers were identified as elements of TATA, perhaps leading to the exclusion of NDM3.

### 3.3.2 Discontinuous Motifs

The split-motif algorithm is especially tuned to discover pairs of TFBS that localize relative to one another. As anticipated, after running the algorithm on *Drosophila* promoters, we found many such pairs. The spacing of the found split-motifs was sometimes in excess of 30bp, indicating true long-range structure in these promoters.

Previous studies have shown that some pairs of TFBS preferentially co-occur by utilizing the position weight matrices (PWMs) of known TFBS and calculating whether any pair of these PWMs occur together greater than expectation. However, these studies have never suggested whether or not there is also a preferential spacing between these motifs, something that could be relevant at a molecular binding level. Secondly, since these studies have utilized the PWMs of known TFBS they have never searched for novel elements. It is possible that a protein complex binds to a long discontinuous DNA motif, each half of which only rarely occurs independently and thus may not be found by traditional means. Because we investigated all split-8mers, our results are not in any way biased toward known TFBS and the algorithm has the potential to discover novel "split" sequence elements.

Some pairs we found that are known to co-localize include DMv4 and DMv5, TATA and INR, and INR and DPE. Even though these are "known" pairs, the spacings between each set has never been identified. However, with the "split-nmer" algorithm that information is built-in. For example, for DMv4 and DMv5 we now know, from the split-motif algorithm, that DMv5 preceeds DMv4 at exactly a

distance of 25 (from the "TGGT" to the "GGTC").

In addition to their preference to co-localize, the preferred position of TATA, INR and DPE relative to the transcriptional start site in a promoter is known down to a base-pair precision. These two facts only imply, but do not prove, that the co-localization and the preferential placement are simultaneous. The split-motif algorithm, however, does prove the connection between co-occurrence and preferential spacing. For INR and DPE/DPE1 this spacing is 27 (from "CAGT" to "CGGT"/"CGGA"). For TATA and INR we found two separate preferred spacings (from the start of "TATA" to the start of "CAGT"): 32 and 33. The canonical 33 spacing correlated with a larger connected component whose edges were more significant. Their were minor differences between the two discovered split-motifs, suggesting that this change in separation may affect how the transcription factors bind to each half of the split-motif.

We also found other pairs of TFBS which had not been previously identified to co-localize, including some novel elements. For example, our algorithm revealed that the DRE and E-box preferentially localize relative each other with a spacing of 25 (from "TCGA" to "AGCT"). The order of these two elements is interchangeable since both motifs approximate palindromes. Both INR and DRE were found to preferentially localize relative to several novel elements. These novel elements are all relatively short (6bp or less), which may be one potential reason they have not been identified as promoter elements in the past. The elements which preferentially localize relative to INR also preferentially localize in the proximal promoter (Figure 3.5). This makes sense since the INR is very well localized in the promoter.

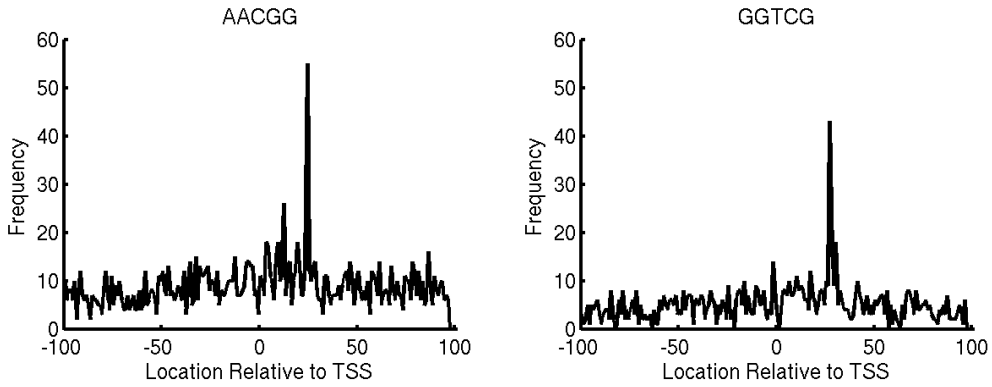



Figure 3.5: The frequency distribution for two sequences found to preferentially localize relative to the INR. These sequences also localize in the proximal promoter.

One of the most interesting pairs of TBFS discovered by the split-motif algorithm was that of DRE-DRE. Pairs of DREs accounted for just over a third of our discovered components and had a spacing length which varied from nearly-overlapping to 35. One potential weakness of the split-motif algorithm is that it utilizes a fixed spacing in the split-8mers when calculating the p-values used to create the correlation network. This might prevent the algorithm from discovering pairs of TFBS which like to co-localize but which are not preferentially localized relative to one another. However, this was not the case for pairs of DREs. Rather than missing the pair, the split-motif algorithm instead found the pair at many spacings, showing that the two halves preferentially co-occur within 35 base-pairs of each other but do not preferentially localize relative to one another within this separation.

## 3.4 Discussion

The novel feature of the split-motif technique is that rather than looking for the co-occurrences of two distinct, known, TFBS, represeted as PWMs, we have looked for the occurrence of a single long motif whose central region is highly degenerate. This allows us to find novel pairs of elements in a computationally efficient manner. Furthermore, since the algorithm joins together many shorter sequences to form a longer motif rather than merely hunting for longer sequences, the results are statistically much more reliable as they are the accumulation of many samplings coming together rather than a single sampling. Because of this we can confidently find motifs which as a whole occur relatively few times but whose individual components all statistically co-occur above the DNA sequence background.

One potential drawback to the algorithm is that it depends on the spacing between the individual "motifs" being fairly consistent in the majority of the occurrences of the split-motif. If two TFBS co-occur but do not co-localize it will be more difficult for our algorithm to link these TFBS. However, as long as the pair co-occurs more often than the background of the individual split-8mers making it up, then the algorithm should be able to find it, as in the case of pairs of DREs.

Since many proteins are fairly well conserved across species, it has been hypothesized that one fundamental difference between species is not in their genetic make-up so much as in how their genes are regulated. Because of this, the properties of promoters are very different in different species. As a consequence, just because the split-motif algorithm was able to discover novel elements in *Drosophila* does not

mean it should work in other organisms. Many motif-finding algorithms work well on prokaryotic organisms only to fail when applied to eukaryotes [54]. However, since the split-motif algorithm is specifically designed to uncover instances of complex regulation, something which becomes more abundant in more complex organisms, it is the authors' hope that it will actually perform better across species than many current novel motif-finding algorithms which focus on single and short TFBS.

## 3.5 Conclusion

Using the split-motif algorithm, we found both novel, continuous TFBS, and also many pairs of TFBS which preferentially localize relative to each another. We looked at the localization of these novel elements and found that many of them also cluster in the proximal promoter, a trait that has been shown to be correlated with a DNA sequence's function. The algorithm's strength is in that it can systematically search for both continuous and discontinuous, or pairs of, DNA sequences and provides a simple way in which to group the discovered elements, a feature lacking in many novel motif-discovery algorithms. It is also able to find these longer and discontinuous motifs with greater statistical confidence. Finally, it assigns a specific spacing between pairs of discovered motifs, something which has never been done before. This preferential spacing may be able to give profound insights how complexes of proteins bind to promoter DNA and act in concert to control gene expression.

## 3.6 Acknowledgements

This work was done in close collaboration with Peter C. FitzGerald at the National Cancer Institute, National Institutes of Health and was in part inspired by the work of Andrey Shlyakhtenko. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster (http://biowulf.nih.gov) at the National Institutes of Health, Bethesda, Md.

## 3.7 Methods

### 3.7.1 Treatment of the input sequences prior to analysis

*Drosophila* promoter sequences for all identified mRNA coding genes were downloaded using the UCSC Genome browser. We ignored annotations for alternate splice sites ("long NMs"). We also removed genes with questionable annotations (genes whose transcriptional start site was annotated within 30bp of its coding start site) and genes whose annotated TSS was too close to the end of the chromosome to contain a full promoter region, defined as from -1000 to +500 around TSS. In order to prevent unintentional sequence duplication, for the remaining genes we also determined the location of transcriptional start sites along the chromosomes and if any pair of genes had an annotated transcriptional start site within 500bp of one another, the latter promoter region was removed from our analysis. This left us with 9,494 *Drosophila* promoters. In this analysis we focused on the proximal promoter, using regions of DNA sequence from -250 to +50 around the TSS. We masked these

sequences using both the standard UCSC repeat masker and a low-complexity filter. For the low complexity filter we used the "dust" program and set the sequence length variable equal to 11.

### 3.7.2 Quickly finding the location of all split-8mers

In order to quickly find the location of split-8mers in an input data set we used the open-source command-line program, TACG, originally designed for restriction enzyme analysis. We began by determining the locations of all 4bp DNA sequences in our input sequences. We then used those locations to construct information regarding the locations of each split-8mer.

### 3.7.3 Visualization of correlation networks

Visualization of our correlation networks was done using the neato program within the Graphviz suite (www.graphviz.org). The color defines the connected components or communities of split-8mers and the length of each edge correlates with the p-value associated with that pair of split-8mers, with the shorter edges representing more significant p-values. We only show links in the correlation network that belong to components containing at least a sixteen members.

### 3.7.4 Sequence alignment

Pairwise and multiple sequence alignments were done using ClustalW. Since our pairs of split-8mers most likely represent overlapping sequences, we set the gap

parameter to 100 to prevent any gaps from opening in aligning the sequences. Once the sequences were aligned, their PWMs were visualized using weblogo by Berkeley. We included the small sample correction option since many times the ends of the aligned sequences are only represented by one or two samples out of the entire set.

### 3.7.5 Functional enrichment analysis

To perform our functional enrichment analysis we used the web-based program DAVID and it's stand-alone counterpart, EASE.

# Chapter 4

# Understanding and Improving Gene Network Reconstruction using Functional Relationships between Genes

If one gene regulates another, those two genes are likely to be involved in many of the same biological functions. With this in mind, we propose a method to create a gene interaction network entirely based on functional annotations within the Gene Ontology (GO). We apply our method to *E. Coli* and find that the strength of links in our ontology-based network is highly correlated with the existence of known regulatory interactions published in RegulonDB. Further, we observe that these strengths are indicative of the importance of links in the known regulatory network's structure. Our ontology-based network is almost as predictive as methods that use gene expression data to calculate mutual information between genes (in particular, we compare our approach to the widely cited context-likelihood-of-relatedness (CLR) algorithm). In addition, the ontology-based approach identifies a different subset of regulatory interactions compared to the mutual information approach. We show that combining predictions from the ontology-based network with those predicted by other reconstruction algorithms leads to a significant improvement in the accuracy of the reconstructed network.

## 4.1 Introduction

The Gene Ontology (GO) [2] provides a controlled setting in which biologists can annotate genes with their functional properties. Since its inception, GO has been applied in various ways, including functional analysis of sets of genes [40] and further annotation prediction [45].

By linking genes based on shared functional annotations, it is possible to construct a gene network with links representing the functional similarity between pairs of genes. However, determining exactly how to calculate this "functional similarity" is non-trivial. We propose a natural weighting scheme under which the "functional similarity" of two genes is correlated with their likelihood to appear as a regulatory interaction in experimental networks. We believe that this model could be used to help produce approximate gene networks for species that do not have a well established gene regulatory network from experiments. In addition, constructing the ontology-based network allows us to interpret the functional role that inferred links might play in the true regulatory network.

In order to evaluate the predictive power of our approach, we compare our ontology-based network with regulatory networks predicted by application to gene expression data of the well-established, context-likelihood-of-relatedness (CLR) network reconstruction algorithm [28]. The CLR algorithm produces a gene interaction network by calculating the mutual information between gene expression data for genes pairs, and by using this in a criterion for judging whether a network link exists between any given pair of genes. We show that our ontology-based network

predicts a biologically distinct subset of regulatory interactions from CLR. This suggests that combining predictions from the ontology-based network with those predicted by gene-expression based reconstruction algorithms might enhance one's ability to reconstruct regulatory networks. By tests using the experimentally determined RegulonDB transcription network for *E.coli*, we verify that this is indeed the case. An outline of this process is shown schematically in Figure 4.1.

In addition to demonstrating that the strength of the links in our ontology-based network is correlated with the existence of a regulatory link, we also find that links which reflect strong connections in our ontology-based network are likely to be structurally important in terms of information flow in the true regulatory network.

We will focus on *E.coli*, since it has been used extensively in training network reconstruction algorithms, and there is a high quality experimental *E.coli* gene network published by RegulonDB [31].

## 4.2 Background

### 4.2.1 Annotation properties of the Gene Ontology

The Gene Ontology takes the form of directed acyclic graph (DAG) with three independent branches: "Molecular Function," "Biological Process," and "Cellular Component." Within each of these branches, genes are annotated to "terms" representing their physical and functional roles in the cell. Terms are organized hierarchically. E.g., a term broadly describing a class of functions may be the "parent" of several "child" terms associated with functions in the broad class of the parent

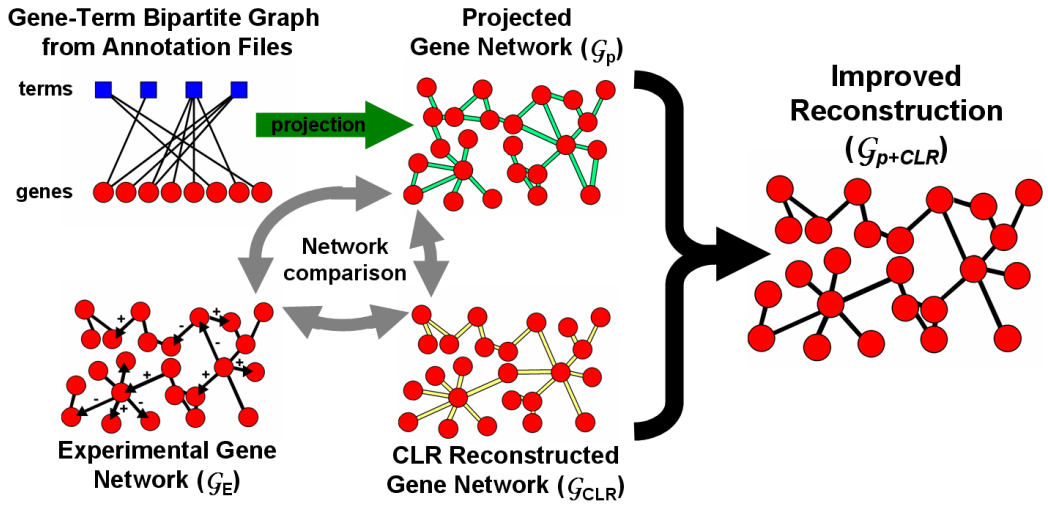



Figure 4.1: Outline of the process. We use Gene Ontology annotations to create a projected gene network ($\mathcal{G}_p$). We then compare this projected network to the experimentally verified network, $\mathcal{G}_E$, published by RegulonDB, to determine which biological regulations our projected network best recapitulates. We also compare $\mathcal{G}_p$ and $\mathcal{G}_E$ to the gene network, $\mathcal{G}_{CLR}$ predicted by the CLR reconstruction algorithm. We find that our projected network predicts a different subset of regulatory interactions than the network reconstructed from gene expression data using the CLR algorithm. We propose combining the results of the ontology-based approach with the gene-expression/mutual-information approach in order to obtain an improved network reconstruction, $\mathcal{G}_{p+CLR}$.

term, and these child terms may be the parents of still more specific terms. Because the gene ontology is a DAG, child terms can have more than one parent term. Gene annotations are transitive up the DAG, meaning an annotation to a child term implies annotations to all the parent terms of that child [84]. As a consequence, all genes will contain an annotation to one or more of the three main branches of the DAG.

In order to construct our ontology-based gene interaction network, we used pairs of gene-term annotations downloaded from the Gene Ontology website (geneontology.org) to first construct a bipartite gene-term network, represented as an $n_T \times n_G$ adjacency matrix, where $n_T$ is the total number of terms and $n_G$ is the number of genes listed in the annotation file. In this matrix a value of one indicates a known connection between the corresponding gene and term, and a value of zero indicates that the gene is not associated with that term. We will denote the $n_T \times n_G$ adjacency matrix of this bipartite graph by $B$ and its $n_G \times n_T$ transpose by $B'$.

Many terms are only associated with a small handful of genes, while some terms are associated with many genes. A histogram of the "degree" of terms in *E.coli* (i.e., the number of genes annotated to each term) reveals a roughly power-law relationship (Figure 4.2). Although there are several different phenomena that could result in a term having a large number of genes annotated to it, in the majority of cases a large number of annotations merely indicates that the functional term is very general and is at the top levels of the DAG. We will exploit this fact when determining the strengths of the functional links between genes in our gene network.

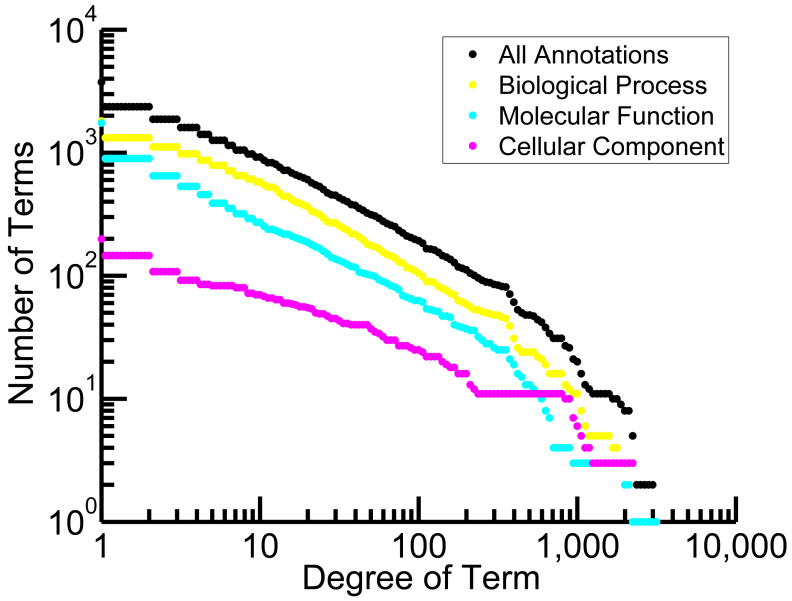



Figure 4.2: Cumulative degree distribution for terms considering all gene-term annotations and just those in each individual ontology.

Each of the three main ontologies also has unique properties. For example, although the "Cellular Component" ontology contains the fewest total terms, and its terms also on average have the highest degree. Therefore, if we use the degree of terms to weight down links between genes, we can expect that our final weights will, on average, have the smallest contribution from "Cellular Component" and the largest contribution from "Molecular Function." This will be discussed in more detail in Section 4.4.2.

### 4.2.2 Expression-based regulatory network reconstruction

There are several popular approaches which utilize high-throughput gene expression data to reverse engineer a regulatory network and create what is known as a relevance network [15]. Many relevance network approaches calculate the corre-

lation or mutual information of the expression profiles for pairs of genes, then link gene pairs with high correlation or mutual information values to build a regulatory network. Recent developments in these computational algorithms have made constructing large and complex networks more feasible and are reaching a point where networks for many systems can be fully computed on a normal desktop computer [55]. However, despite major advances and evidence of their ability to lend biological insight [6], the quality and completeness of these biological networks still needs improvement. Comparisons between networks generated using these algorithms and known regulatory networks, such as the one put forth for *E.coli* by RegulonDB [31], shows only about a 60% ability to predict true regulatory interactions [28], and this is only for a specific subset of all known regulatory interactions.

Some of the most successful algorithms for generating gene regulatory networks involve the information theory concept of mutual information (MI) [94][55][28]. MI describes the statistical dependence between two variables. However, unlike traditional correlation coefficients, MI does not assume a linear relationship and has been used to successfully detect regulatory interactions which would have been missed using a linear correlation metric. The MI between genes $a$ and $b$ is:

$$MI(a,b) = \int log \frac{p(\alpha,\beta)}{p(\alpha)p(\beta)} \, d\alpha d\beta, \tag{4.1}$$

where $\alpha$ and $\beta$ denote the expression levels of genes $a$ and $b$, and $p(\alpha,\beta)$, $p(\alpha)$ and $p(\beta)$ are joint and marginal probability densities of these expression levels. In practice, several different measurements $(\alpha_i, \beta_i)$ are obtained, possibly under

different perturbing conditions, and used to obtain estimates of $MI(a, b)$.

We wish to investigate how the results of a reconstruction approach using the publicly curated functional data in GO compares to results of techniques using gene expression data such as the ones described above. As an example, we will compare our ontology network to the network produced using the context-likelihood-of-relatedness (CLR) algorithm for *E.coli*. CLR calculates the MI between each pair of genes. It then performs a background correction step in order to eliminate false correlations by calculating a "Z-score" defined as:

$$Z(a,b) = \sqrt{\left(\frac{MI(a,b) - \mu_a}{\sigma_a}\right)^2 + \left(\frac{MI(a,b) - \mu_b}{\sigma_b}\right)^2}, \tag{4.2}$$

where $\mu_x$ and $\sigma_x$ are the average and standard deviation of the MI values associated with gene $x$, respectively. If the Z-score value is above a chosen threshold, a network link is judged to connect genes $a$ and $b$. We explore whether functional relationships are reflected in networks reconstructed by the CLR algorithm and whether the types of interactions captured by the ontology-based network are fundamentally different from the types captured by the CLR network.

## 4.3 Approach: Gene Networks based on Gene Ontology

From the adjacency matrix $B$ of our bipartite graph we can generate an adjacency matrix $\hat{G}$ specifying a projected network relating genes annotated in the Gene Ontology:

$$\hat{G} = B'B, \qquad \hat{G}_{ij} = \sum_{p} B_{ip} B'_{pj}. \tag{4.3}$$

In this projection the value of $\hat{G}_{ij}$ is equal to the total number of functional annotations which are shared between genes $i$ and $j$. However, as previously discussed, some terms such as "Molecular Function" are quite general and associated with many genes, while others are only associated with very few genes. It would, therefore, seem inappropriate to weight links between genes $i$ and $j$ simply by the number of their co-associations with terms (as done in the above definition of $\hat{G}_{ij}$). E.g., one might want to count associations through terms that have many gene annotations less strongly those associations that occur through more specific terms (like "beta-catenin binding," which has only a handful of gene annotations). One common practice employed to address this issue is to ignore the highest level of the DAG [50]. However, we choose instead to compensate for the variation in the quantity of term annotations by introducing a diagonal weighting matrix which is based on the degree of terms in $B$:

$$w_{ij}^{(\alpha)} = \frac{\delta_{ij}}{\left(\sum_{p=1}^{n_T} B_{pi}\right)^{\alpha}}, \tag{4.4}$$

where $\delta_{ij} = 1$ if $i = j$ and is zero otherwise. Using the matrix $w^{(\alpha)}$, we modify the strength of our gene connections as given in Equation 4.3 to obtain a new gene conntection matrix $\hat{G}^{(\alpha)}$ given by:

$$\hat{G}^{(\alpha)} = B' w^{(\alpha)} B, \qquad \hat{G}^{(\alpha)}_{ij} = \sum_p \frac{B'_{ip} B_{pj}}{(\sum_l B_{pl})^\alpha}. \tag{4.5}$$

where $\alpha$ can be thought of as a weighting parameter such that larger $\alpha$ more strongly suppresses the weights of terms (index $p$ in Equation 4.5) that have connections to many different genes (index $l$ in Equation 4.5). Note that for $\alpha = 0$, the weighting matrix, $w^{(\alpha)}$ reduces to the identity matrix (i.e., uniform weighting of terms), and $\hat{G}^{(0)} = \hat{G}$. Another issue is that, if two genes are annotated to the same term, those two genes will also both be annotated to all the parents of that term. However, these annotations are redundant. Therefore, to compensate for this we further modify our projected gene-gene adjacency matrix of Equation 4.5, as follows:

$$G^{(\alpha)}_{ij} = \sum_p \frac{B'_{ip} B_{pj}}{(\sum_l B_{pl})^\alpha} \delta_{ipj}, \tag{4.6}$$

where $\delta_{ipj}$ has been introduced to compensate for the above-described redundancy of annotation. There are two choices for $\delta_{ipj}$ which we employ.

*(i) Lowest-Level-Annotation choice*: One option is to take only lowest-level annotations in the Gene Ontology, which are, by their nature, non-redundant. In this case $\delta_{ipj}$ will equal 1 if gene $i$ and gene $j$ both have a lowest-level annotation to term $p$, and zero otherwise. One potential weakness of this method is that only genes which share a lowest-level annotation will be linked, and thus our network may be relatively sparse.

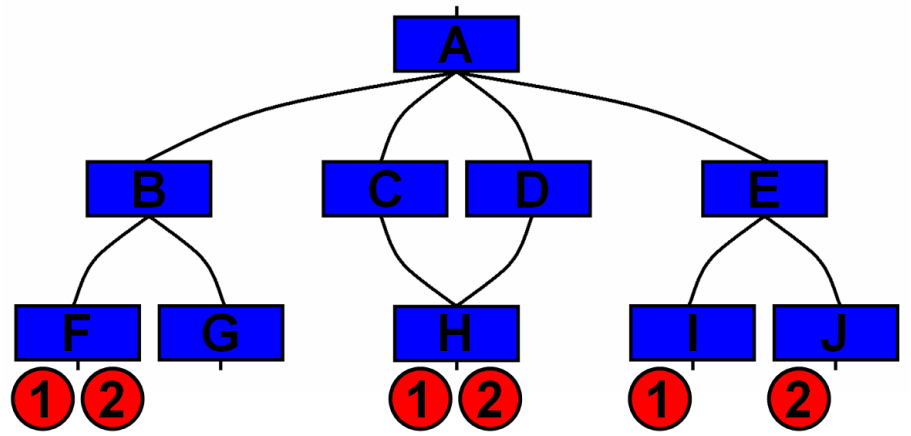
*(ii) Lowest-Common-Ancestor choice*: Another viable alternative is to consider

all lowest-common annotations between pairs of genes, meaning that if two genes are both annotated to the same term, consider that common annotation, but none of the annotations to the parents of that term. In this case $\delta_{ipj}$ will equal 1 if term $p$ is a lowest-common-ancestor of gene $i$ and gene $j$, and zero otherwise.

Figure 4.3 and its caption illustrate these two choices for $\delta_{ipj}$. We note that use of $G^{(\alpha)}$ in place of $\hat{G}$ turns out to be crucial in out network construction, and we will investigate how varying the weighting parameter, $\alpha$, affects the predictive power of our method. For ease of notation we will henceforth drop the subscript, $(\alpha)$, on $G^{(\alpha)}$. That is, we use $G$ to denote $G^{(\alpha)}$.

Since the Gene Ontology has three distinct branches, we will also investigate the effects of considering the individual ontologies of each of these branches as opposed to the entire gene ontology. Therefore, we will have four versions of $G$: (i) the reconstruction considering all gene-term annotations ($G^{All}$), (ii) considering only gene-term annotations where the term is part of the "Biological Process" ontology ($G^{BP}$), (iii) considering only gene-term annotations where the term is part of the "Molecular Function" ontology ($G^{MF}$), and (iv) considering only gene-term anotations where the term is part of the "Cellular Component" ontology ($G^{CC}$). Because $G$ in Equation 4.6 is defined as a sum over terms (i.e., the index $p$ in Equation 4.6),

$$G^{All} = G^{BP} + G^{MF} + G^{CC}.$$




**All Common Annotations of (1) and (2): A, B, C, D, E, F, H**

**Lowest-Common-Ancestors of (1) and (2): E, F, H**

**Lowest-Level-Annotations of (1) and (2): F, H**

Figure 4.3: An illustration of the difference between the lowest-common-ancestors choice of $\delta_{ipj}$ and the lowest-level-annotations choice of $\delta_{ipj}$. If gene (1) and gene (2) are annotated to terms F, H, I, and J, as illustrated, they will share annotations to terms A, B, C, D and E. However, the shared annotation to terms A, B, C and D are redundant in that they are a consequence of the shared annotations to terms E, F and H. E, F, and H are the "lowest-common-ancestors" of genes (1) and (2), since the shared annotation between gene (1) and (2) through these terms cannot be attributed to a shared annotation at a lower level of the hierarchy. In the "lowest-level-annotation" weighting scheme we take this one step further and only consider annotations between two genes which are not a consequence of annotations lower in the Gene Ontology DAG. In the illustrated example, this would be analogous to only considering annotations to terms F and H, and not E.

## 4.4 Results

### 4.4.1 The effects of weighting $G$

There are several limiting cases for $\alpha$. For $\alpha = 0$ the weighting matrix reduces to the identity matrix and the calculation is the same as it would have been had we not considered any weighting. In this case the values of $G$ will be the number of terms shared between two genes ($G = \hat{G}$). For large $\alpha$ the weights of $G$ are such that those genes connected through many low degree terms have the highest weight and those connected through only one high degree term have the lowest weight. Low-degree terms (i.e., terms with few gene annotations) are normally lower in the DAG hierarchy and in general represent more specific biological functions. Therefore, by giving the greatest weight to links between genes which share annotations to many low-degree terms, our weights should correspond to a measure of how much specific biological function is in common between the two genes.

To determine the consequences of different weighting parameter values, we compared our projected network for *E.coli* for various values of $\alpha$ to the established regulatory network published by RegulonDB. RegulonDB provides a high-quality TF-gene interaction network which contains 1987 genes and 5717 regulatory links. Of these 1987 RegulonDB-listed genes, 1729 also appear in the GO annotation files, and of the 5717 RegulonDB-listed links, 1408 also appear in our lowest-level-annotation projected gene network, and 4815 also appear in our lowest-common-ancestor projected gene network. We believe that this provides sufficient shared information for us to usefully compare our projected gene networks with the

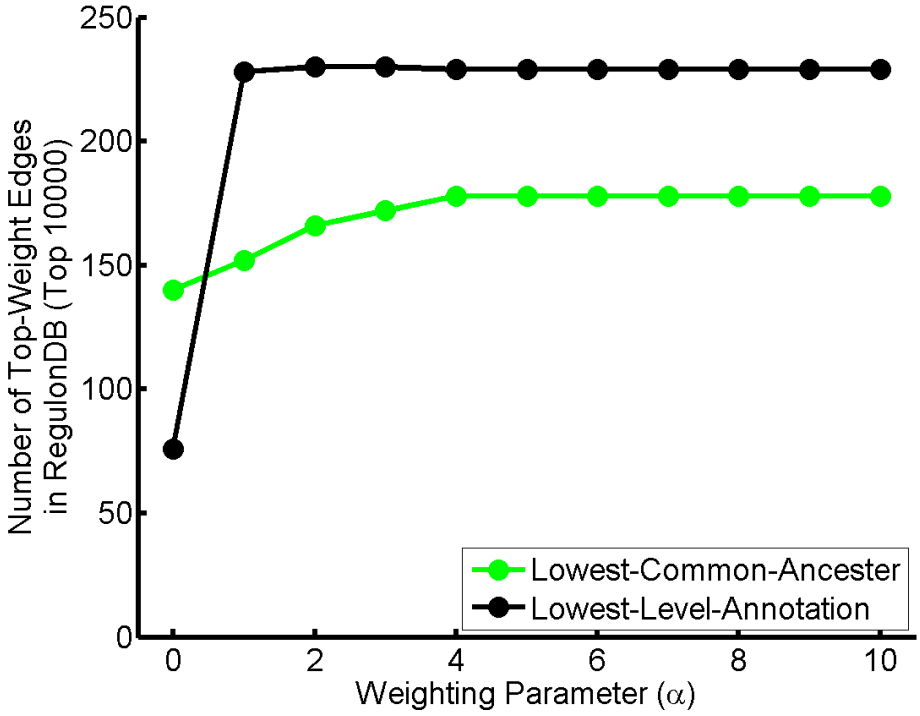



Figure 4.4: A comparison of the two proposed weighting schemes and the effects of the weighting parameter $\alpha$ on the predictive power of the projected networks.

experimentally-derived RegulonDB gene network.

Figure 4.4 shows a comparison of the predictive power of the lowest-level-annotation and the lowest-common-ancestor weighting schemes as a function of the weighting parameter $\alpha$. As expected, for both weighting schemes the results for high $\alpha$ reach a steady value. Although the network is much sparser, the weighted values using only lowest-level annotations are more predictive of true regulatory interactions than those calculated using all lowest-common-ancestors. Note that for those edges which appear in both weighting schemes, the only differece in weight is the addition of shared annotations which are lowest-common-ancestors but not lowest-level-annotations. Additional edges which do not appear in the lowest-level-

annotation weighting scheme are links between genes which do not share a single lowest-level-annotation. Note that the lowest-common-ancestor annotations, even when through very specific, low-level terms, are apt to produce false positives and reduce the predictive power of the edges.

We now discuss the issue of whether a link between genes in the ontology-based network implies that one gene regulates another or that they are both regulated by a third gene. When we consider lowest-level annotations, we are only considering pairs of genes which are known to both perform the same specific function. On the other hand, when considering lowest-common-ancestor annotations, we allow the link between two genes to be strengthened if those genes perform many similar, but less specific functions. In the limit of large $\alpha$ this weighting converges to the inverse of the degree of the highest degree term linking the two genes, independent of how specific that function is to the behavior of each of the individual genes. Therefore, it is reasonable to suppose that lowest-level-annotations are more likely to represent direct interactions between pairs of genes, whereas functions which are the lowest-common-ancestor of two genes may represent more indirect interactions. This is consistent with the enhanced performance of the lowest-level-annotation model over the lowest-common-ancestor model. Because of this enhanced performance, we will show results only for the lowest-level-annotation model in the following sections.

### 4.4.2 The role of the three ontologies in the weighting of the projected gene network

In addition to applying our approach to the entire GO hierarchy, we also used it (with the lowest-level-annotation weighting scheme) to determine separate gene interaction networks for each of the three main branches of the GO hierarchy. As expected, the total weight contribution to the edges in the composite ontology-based network is reflective of the degree-distribution of the terms within each of the three individual ontologies (see Section 4.2.1). 35% of the total edge weight came from "Biological Process," 54% from "Molecular Function," and 11% from "Cellular Component."

It is informative to look at how each of the three main ontologies contributes to individual edges within our network. Each edge can be broken into the weight contributed from each of the three main ontologies, and the percentage of each of these ontology's contribution to that edge's weight can be calculated. Three-quarters of the edges in our projected network have weights determined by annotations in only one of the three ontologies, 11% from only "Biological Process," 38% from only "Molecular Function," and 26% from only "Cellular Component." Of the remaining 25% of the edges, 9% have their largest contribution from "Biological Process," 13% have their largest contribution from "Molecular Function" and only 3% have their largest contribution from "Cellular Component."

Although a smaller percentage of edges have weights dominated by the "Biological Process" ontology compared to the other two ontologies, these edges are

highly over-represented in the top weighted edges in our projected network. For those edges whose contribution is from "Biological Process" alone, 92% appear in the top 10,000 weighted edges. For edges whose weight is most influenced by "Biological Process," 95% appear in the top 10,000 edges. In fact, over half of the top weight edges are dominated by annotations from the "Biological Process" ontology.

### 4.4.3 Comparison to other network reconstructions

In order to determine the usefulness of our projected gene network compared to other common network reconstruction approaches, we compared our projected gene network to the one determined by the CLR reconstruction algorithm [28]. The full Z-score matrix provided by CLR contains information for 4345 genes, 3523 of which are also considered in our projected network, and 5.6 million interactions ($\sim$53% of all possible interactions). In comparison, our projected network contains information for 3734 genes and 2.1 million interactions ($\sim$30% of all possible interactions). 3959 interactions from RegulonDB have a Z-score and/or an edge weight.

A little over half (1.2 million) of the edges in our projected gene network have corresponding Z-score values determined by CLR. For these edges we determined the rank order of our annotation edge weights and the rank order of the Z-score value. We further identified which of theses edges are also listed as true regulatory interactions by RegulonDB and have illustrated the results in Figure 4.5. Although edges with both high Z-score and high edge weight are most likely to be in RegulonDB, there are very few edges predicted by CLR which are not also predicted by

our method. Furthermore, edges with high edge-weight are more likely to be in RegulonDB even if they have low Z-Scores. Edges which do not have a Z-Score assigned to them by CLR but have a predicted annotation weight show a similar trend. Conversely, for edges with a Z-score but no predicted annotation weight, the Z-score, although predictive, has a higher false-negative rate compared to edges with low annotation weights. This may indicate that edges with no weight are due to a lack of information in the Gene Ontology. Comparing the functional edge weight with the CLR Z-Score illustrates how these two different approaches are likely measuring different biological information.

### 4.4.4 Properties of high annotation weight edges

At this point, is unclear how the weights of edges in our ontology-based network reflect biological information. Clearly, on some level, they represent the functional relatedness between two genes, but what does this property actually mean for the function of the gene regulatory network of an organism? To address this question, we access how information flow changes in the established regulatory network when high annotation weight edges are removed. To systematically explore the connection between annotation weight and information flow, we remove known regulatory links from the experimental regulatory network one at a time and evaluate the change in information flow. In order to measure the change in information flow in the regulatory network we determine the new shortest path between pairs of genes upon the removal of their regulatory link. Edges whose removal causes little difference in

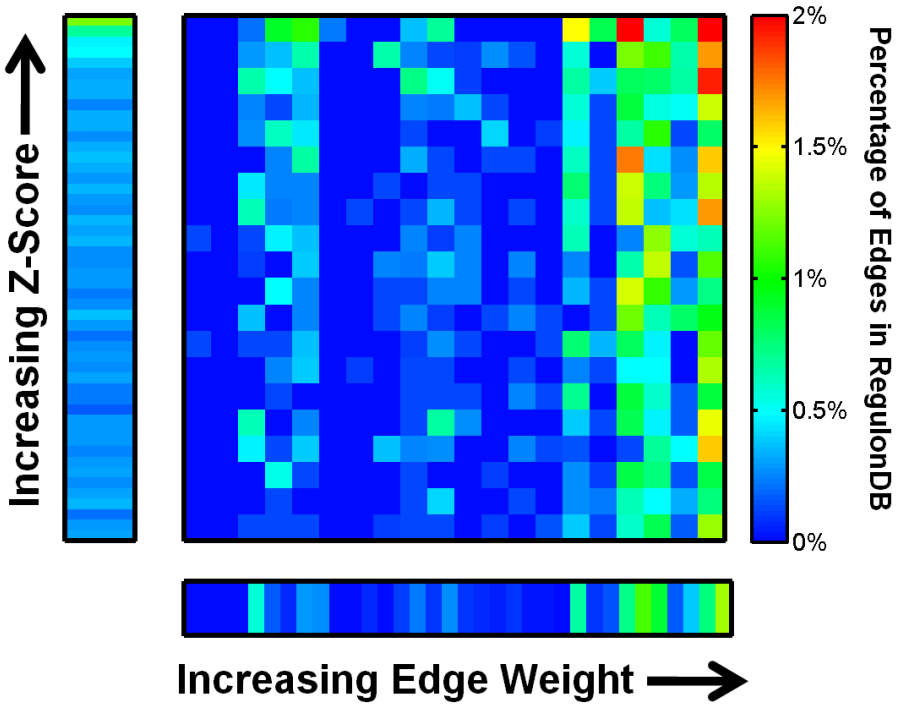



Figure 4.5: Comparison of edge prediction by weighting in our Projected Gene Network versus the Z-Score Gene Network predicted by the CLR algorithm. We ordered the edges which appear in both networks both by their increasing weight and by their Z-Score. We then visually represented which of these edges were also in RegulonDB. This was done by dividing the weight/Z-score plain into 400 (20x20) equal squares with a length and height of approximately 6,000 edges and then determining the percentages of edges in each square which are also in RegulonDB. This percentage is represented as a color in the square. Edges which have only a Z-score in CLR, or only a weight in our projected network, were ordered separately and their predictive ability is illustrated as color-bars along their corresponding axes. These color-bars were created by dividing the ordered edges into equal-sized bins of approximately 6,000 edges (for consistency with the 2-dimensional data), and calculating the percentage of edges in each bin which are also in RegulonDB. This percentage is represented as a color-strip in the color-bar.

the length of shortest path between the genes that it connects can be thought of as redundant, since the two genes are still closely connected in the regulatory network and the edge removal only has minimal effect on the network flow. On the other hand, edges whose removal causes the regulatory path between the two connected genes to increase substantially, or even dissapear, are *informationally important* in the regulatory network.

In order to display whether high annotation weight edges tend to be either redundant or essential to information flow, we ordered the edges in RegulonDB according to their weight in the projected gene network and then calculated the harmonic mean of the new shortest path for edges at or above each indexed value. The results are striking. Those edges with the highest weight are highly important to information flow (Figure 4.6). If we instead rank edges in RegulonDB based on their Z-score in the CLR algorithm, there is only a slight preference for essential edges to have high Z-scores.

This result is curious since one might suspect that genes which share many low-level annotations may be in locally dense regions of the regulatory network and hence exhibit redundancy. On the other hand, the fact that functionally similar genes compose the links which are more important for information flow may actually be indicative of these multiple regulatory pathways flowing through that pair of genes. These pathways may connect communities of genes which are independently involved in only a subset of functions, but which at times must be combined to perform higher-order biological tasks.

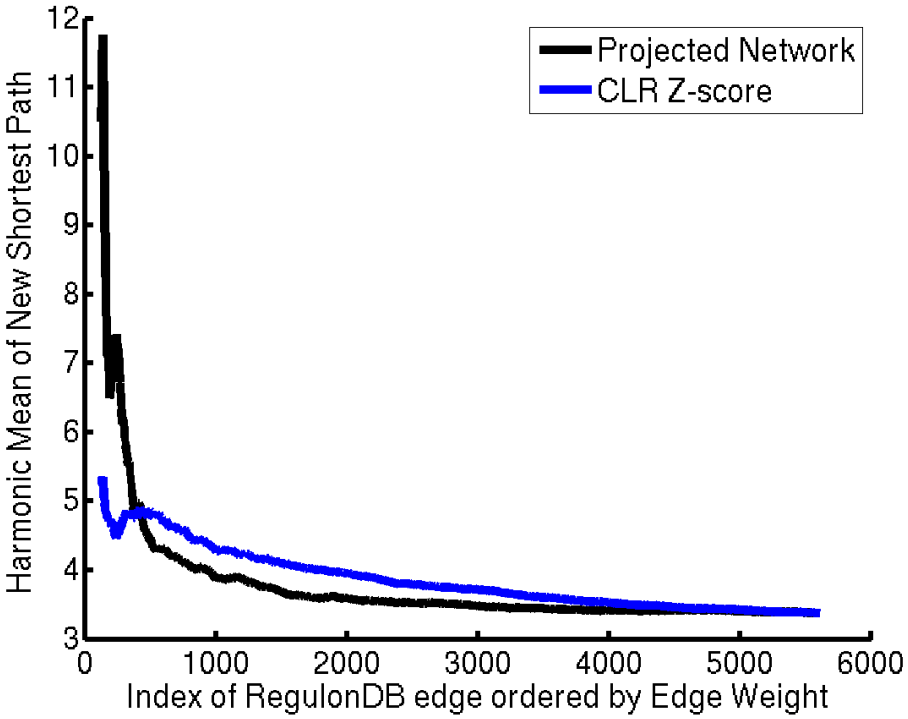



Figure 4.6: A plot of the harmonic mean of the new shortest path in the published regulatory network from RegulonDB as a function of either the edge weight in our projected network or the Z-score in the CLR network. Edges with high weights in our projected network appear to be important for information flow through the published regulatory network. On the other hand, edges with high Z-score, although more informationally important than average, do not have so clear a signature. Note that edges in RegulonDB which do not appear in our projected gene network were given a weight of zero for index ordering purposes. Similarly edges in RegulonDB which do not appear in the CLR Z-score matrix were given a Z-score of zero for index ordering purposes. As a consequence, only edges indexed below 1400 for our projected network and edges indexed below 3535 for the CLR network have weight/Z-score value in their corresponding gene network.

### 4.4.5 Combining GO to improve reconstruction

In order to determine the predictive power of their algorithm for identifying links involving transcription factors, Faith et. al. used a "putative set" of edges in their network, which included any edges connected to a particular set of 328 transcription factors. Within this set of edges their algorithm's performance was dramatically enhanced. In order to better compare the predictive power of our projected gene network with the CLR algorithm, we took the same set of 328 transcription factors identified by Faith et. al. and looked at the subset of edges in our ontology-based gene network that were connected to one of those transcription factors.

In contrast to a comparison including all edges, for which the predictive power of the weights in our ontology-based gene network is equivalent to the predictive power of the Z-score (Figure 4.7(a)), in a comparison that considers the putative set of edges, CLR performs significantly better for the very top indexed edges, but quickly becomes only a mild improvement over our projected network, with only a few percentage increase over our projected network once more than the top 1000 edges are considered (Figure 4.7(b)).

Since CLR and our projected network are identifying a unique subset of edges and both are performing with approximately the same accuracy, we are able to combine them in order to take advantage of the strengths of both algorithms. As a simple model, we consider the same fraction of top edges in both CLR and our projected gene network and determined what percentage of this combined set was

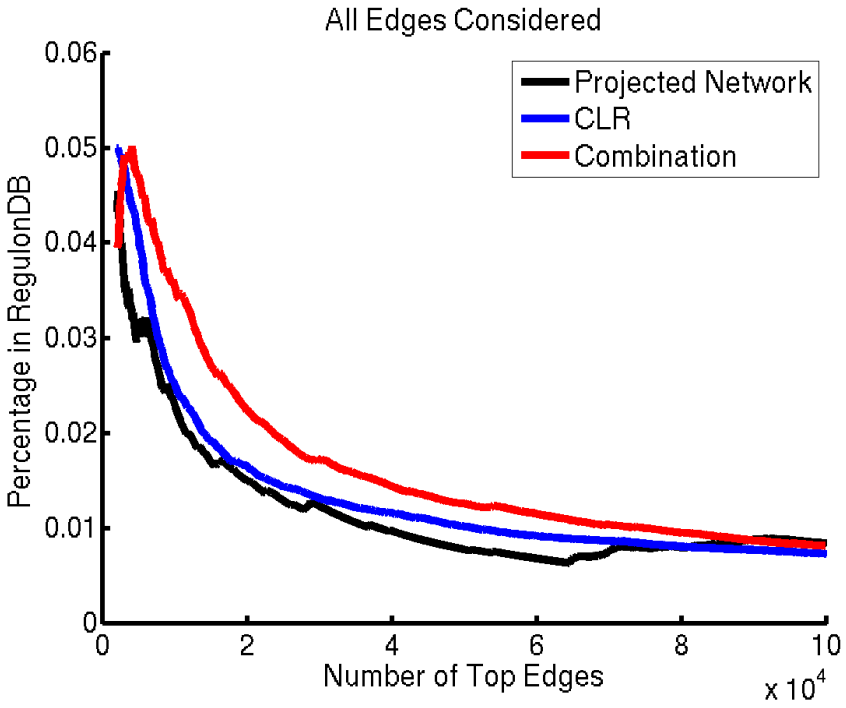
also in RegulonDB. When the size of the combined set reaches more than around 500 edges, this combined model outperform the predictive power of either individual algorithm. We observe a substantial improvement when at least 2000 edges are included by this method (Figure 4.7).

Previous groups have mentioned the power of combining various algorithms to improve network reconstruction [44][50][93]. In this case, the Gene Ontology provides a unique addition to the set of predicted edges since those added are known to be functionally related and, furthermore, as we demonstrated above, they are now also know to be links essential to the flow of information within the regulatory network.

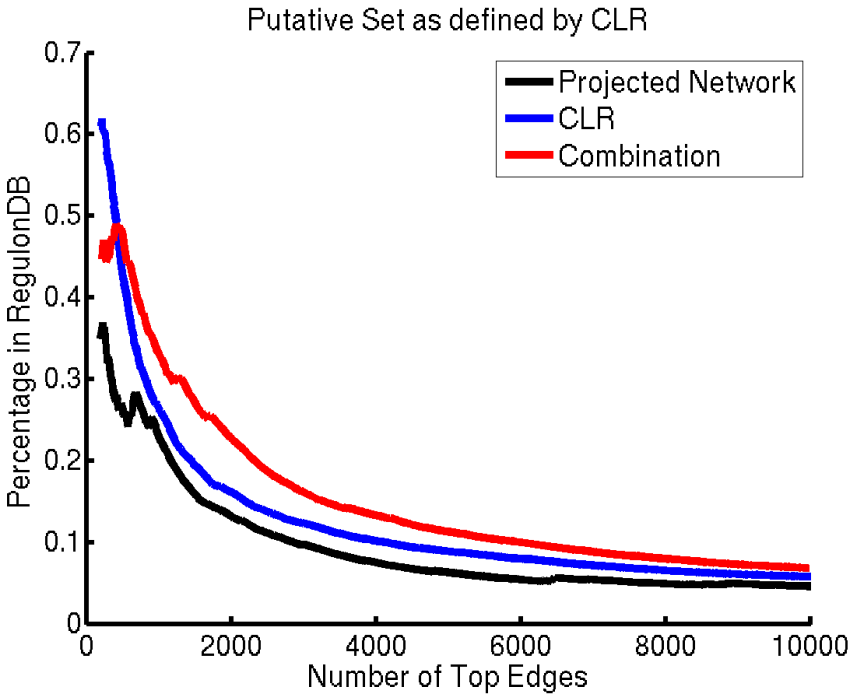
## 4.5 Discussion

Although using the Gene Ontology to create a gene relatory network has many limitations compared to other techniques using other data, the method we describe here is cheap, computationally simple and takes advantage of a large amount of data which has already been accumulated. Whereas algorithms such as CLR tend to focus on one type of biological data, the Gene Ontology contains data from many types of experiments. Furthermore, because it is a publicly accessible collection of data, errors in annotation are likely to be caught and corrected as members of the community use the ontologies for their wide variety of applications.

One of the main weaknesses of using the Gene Ontology to interpret biological data is that there may be subjective biases in its construction. The DAG structure in




(a) Predictive Power of All Edges




(b) Predictive Power in the Putative Set

Figure 4.7: The predictive power of the weights of the projected gene network vs. the Z-score of the CLR algorithm considering both (A) all edges in RegulonDB and (B) the putative set of edges as defined by Faith et. al. Also noted on the images is the predictive power of a combination of the algorithms which considers an equal number of top edges from both algorithms and determines the predictive power of that combined set.

particular, although clearly a reasonable way to classify functional terms, can make it tedious to interpret the meaning of shared annotations between a single pair of genes. Many shared annotations between two genes may be indicative that their functional similarity can easily be classified as a subset of many other functions in a human interpretation of how all functions are related, rather than how important or specific that functional similarity is. Although we believe that our method avoids this issue by eliminating redundant annotations and eventually considering only lowest-level annotations, information may be lost through this elimination.

One advantage of using the Gene Ontology to predict a regulatory network is that biological meaning can more easily be assigned to a predicted interaction via the functional terms used in creating and weighting that interaction. This information in itself is little more than another way to parse data in the Gene Ontology in order to assign a measure of functional similarity between pairs of genes, something that certainly could be done many different ways. However, the fact that the strength of the functional similarity between two genes is correlated with the likelihood for those genes to appear in a regulatory network and, furthermore, is predictive of that link's importance in information flow through the regulatory network, gives much wider range implications on the Gene Ontology's use in not only constructing, but evaluating, regulatory networks.

## 4.6 Conclusion

By using the Gene Ontology to access the interaction between pairs of genes, we are able to give functional meaning to links in a known regulatory network. Although it was previously unclear how to compare the functional similarity of two genes, we demonstrate that, at least within a regulatory network context, lowest-level annotations contain the most information regarding regulatory potential. Even if this similarity is somehow a consequence of human bias in annotation (e.g., it is witnessed that two genes are related in a regulatory fashion therefore they are given a common annotation), it does not explain why genes with the highest weights in our projected network are also the most important for information flow in an established experimental regulatory network. In the future, it will be interesting to test if this striking feature continues to manifest itself in experimental regulatory networks of other species as they become available at increasing quality.

We show that the combination of our functional network reconstruction technique with a gene-expression-MI reconstruction algorithm out-performs the predictive ability of either method alone. Since the predictive abilities of most current reconstruction algorithms are relatively limited, this type of hybrid combined technique appears promising. We believe that using functional data as one input can not only improve network reconstruction, but can also give real biological meaning to the links within this reconstruction. In particular, we hypothesize that links with high functional similarity should be more important to information flow. Finally, using the Gene Ontology in the context of a regulatory network can have applica-

tions to therapeutic techniques. E.g., therapies could be developed which target genes and interactions located in part of the reconstructed network that appear to be associated with biological functions deemed to be relevant to desired therapeutic effects.

# Chapter 5

# Building an Alternate Gene Classification Scheme from Network Structure within the Gene Ontology

The Gene Ontology (GO) provides biologists with a controlled terminology that describes how genes are associated with function and how those functional terms are related to each other. These term-term relationships take the form of a directed acyclic graph (DAG). However, we propose that the graph structure of gene-term annotations found in GO can be used to establish an alternate natural way to group the functional terms which is different from the hierarchical structure established in the DAG. Grouping terms by this alternate scheme provides a new framework with which to describe and predict the functions of experimentally identified groups of genes. In this chapter, we show that gene signatures for cancer that are enriched with respect to branches of the DAG are also enriched with respect to the groups of terms identified through our alternate approach.

## 5.1 Introduction

Although the gene ontology [2][78] has been around for just over a decade, there has only been minimal investigation on how biological functions might be related to each other outside of the established DAG structure, and the majority of this has focused on discovering individual links, especially between ontologies [45],

rather than investigating whether the structure as a whole is the only legitimate way to classify biological functions. We propose an alternate viable way in which to classify functional terms which is dissimilar to the GO hierarchy. This classification scheme relies on the network structure of gene-term annotations rather than pre-established functional relationships and thus represent a completely different type of biological classification.

We begin our study by investigating the annotation properties of the gene ontology as these play a key role in evaluating and using the underlying network structure (Figure 5.1). In the approach section, we describe the method we use to transform the annotation information into term-term relationships in order to create species-specific networks. In the results section we use community structure finding algorithms to partition these term-term networks into communities of terms. We compare the partitions identified to the structure of the GO hierarchy and show that there are strong differences between these two ways of organizing functional terms.

Hence, we propose that our found communities of terms may provide an alternate natural framework with which to use annotations of the gene ontology. In order to verify the biological validity of these term communities, we evaluated the enrichment of cancer signatures in our communities as well as in branches of the hierarchical DAG classification system. We found that cancer signatures were enriched in both our discovered term communities and branches of the GO DAG. We therefore suggest that by linking GO terms based on shared genes, we can create an alternate, biological meaningful, network representation of the functional terms

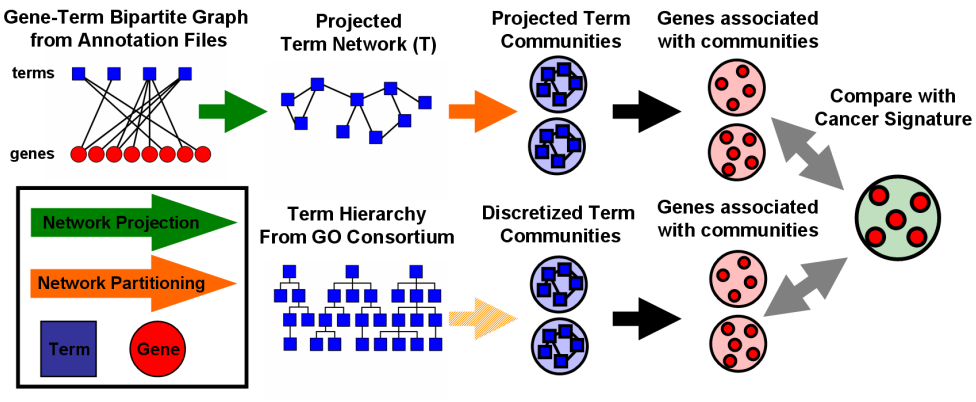



Figure 5.1: Outline of the process. We use Gene Ontology annotations to form a gene-term network. From this network, we can create a projected term-term network. We then partition the term-term network into communities. Simultaneously we will determine communities of terms which represent branches of the Gene Ontology hierarchy. Because the hierarchy takes the form of a DAG, terms can appear in different branches, but we create a discrete partitioning. We compare this discrete partitioning to the one created from our projected term-term network and determine how similar our communities are to the GO hierarchy. Finally, we identify the genes which are annotated to terms in each community, and compare these sets of genes to known cancer signatures to determine the biological viability of our projected communities compared to "known" communities in the DAG.

in GO that is distinct from the established hierarchy.

In this study we will primarily focus on Human annotations so that we can easily access the biological validity of our analysis by evaluating term enrichment in various established cancer signatures. The methods developed in Human can then be used to generate and compare term networks for several additional species, including *E.Coli*, Yeast, *Dicty*, *C.elegan*, *Arabidopsis*, *Drosophila*, Mouse, and Rat. We will compare the communities of terms for each species to evaluate whether a species-independent term-classification scheme is viable or if term relationships vary over evolutionary history.

## 5.2 Background: Annotation Properties of the Gene Ontology

### 5.2.1 GO as a bipartite graph

In the past bipartite graphs have been used to predict and classify biological information [92]. In order to construct our term-term network, we first had to construct a gene-term bipartite graph. This was done from downloaded annotation files that list gene-term pairs (see Section 5.7). The bipartite graph can be mathematically represented as an $n_T \times n_G$ adjacency matrix, where $n_T$ is the total number of terms and $n_G$ is the number of genes listed in the annotation file. In this matrix a value of one would indicate a known connection between the corresponding gene and term, and a value of zero would indicate that the gene is not associated with that term. We will represent the $n_T \times n_G$ bipartite graph adjacency matrix as $B$ and its $n_G \times n_T$ transpose as $B'$.

Many terms are only associated with a handful of genes, however, some terms are associated with many genes. Summing over the rows of $B$ gives the degree of terms (the number of genes annotated to a term) in the bipartite graph. Summing over the columns of $B$ gives the degree of genes (the number of terms to which that gene is annotated) in the bipartite graph. A histogram of the degree of terms in Human reveals a roughly heavy-tailed relationship (Figure 5.2(a)), with a handful of terms containing a large number of annotations and many terms containing only a small number of gene annotations.

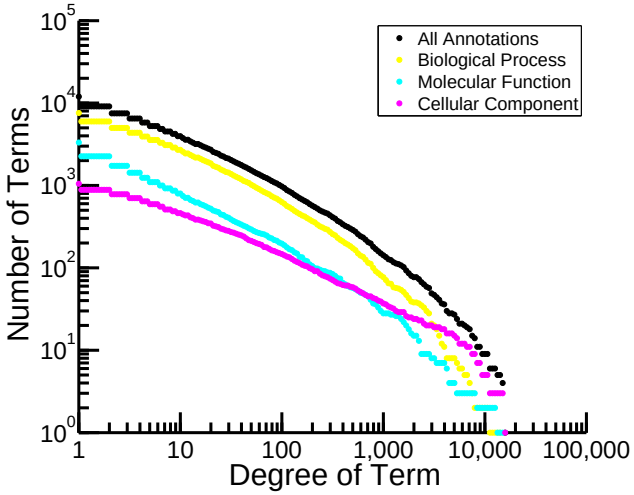
Several different phenomena may cause a term to have a large number of genes annotated to it. There may be many genes which perform that function or the

function may be especially interesting or easy to study. However, in the majority of cases a large number of annotations indicates that a functional term is very general (as in the case of the top levels of the DAG). We will be exploiting this fact when constructing term-term networks based on annotations in the Gene Ontology.

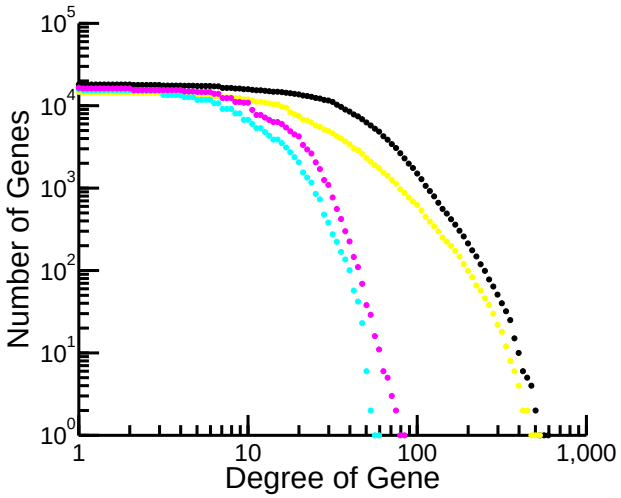
Compared to the term degree distribution, the degree distribution of genes is flat over several orders of magnitude (Figure 5.2(b)). The degree of a gene is generally indicative of how much specific information is known about that gene and is generally correlated with how many low level annotations the gene has rather than the placement of these annotations in the DAG structure.

### 5.2.2 The three main ontologies

The Gene Ontology is a directed acyclic graph (DAG) which has three independent branches: "Biological Process," "Molecular Function," and "Cellular Component." Terms may have multiple parents and multiple children but can only belong to one of the three main ontologies [84]. Because of this there are no defined connections between terms assigned to different ontologies. Nonetheless, there are biological cases in which a term in one branch is related to a term in another branch. It is also probable that many terms within the same ontology are related even though they are not connected in the DAG. We wish to address two main questions. Firstly, whether the established DAG structure is the only natural way in which to classify these functional terms and secondly, if such an alternate classification exists, how could it be used to evalute biologically related sets of genes.




(a) Degree Distribution of Term Annotations




(b) Degree Distribution of Gene Annotations

Figure 5.2: Cumulative degree distribution for (A) terms and (B) genes considering all gene-term annotations and just those in each individual ontology.

The three ontologies themselves have unique graph properties. The number of terms, number of annotations, and the average number of annotations made to each term varies across the ontologies. "Biological Process" is the largest ontology, containing both the most number of functional terms and total gene annotations. Although "Cellular Component" has the fewest number of functional terms and number of annotations, on average, it has the highest number of average annotations per term.

We investigate the effects of considering each of the three individual ontologies separately as opposed to examining the entire gene ontology as a whole. This is done by first projecting the bipartite network of each individual branch to a term-term network for that branch. Then we will consider gene-term annotations for all branches and compare the resulting term-term network to the ones derived for isolated branches.

## 5.3 Approach: Projecting Term Networks based on Gene Ontology

Varying the weights in the gene-term bipartite graph can significantly alter the communities of terms that are identified. An appropriate weighting method is crucial for properly characterizing the graph structure of the Gene Ontology bipartite graph, and hence identifying biologically relevant communities of terms. A comprehensive discussion of various weighting schemes and the consequences of those weightings is discussed in Appendix C.

From our bipartite graph ($B$) we could immediately create a projected term

network by joining together any pair of terms which share common genes. However, this approach would lose a large amount of information. For example, connections between terms with a high degree would be very abundant merely because they are more probable and not because they are more informational. Furthermore, terms with a high degree are for the most part more general, and giving them more weight in our projected network may disguise the most biologically informative community structure. Previously, others have compensated for the disparity in term significance by ignoring the highest levels of the GO DAG [50][39]. However, we instead choose to correct for the skewed term degree distribution by constructing a diagonal weighting matrix, with off-diagonal elements equal to zero and diagonal elements equal to:

$$w_{jj} = \frac{1}{\sum_{i=1}^{n_T} B_{ij}},$$

or simply one over the degree of the term in the bipartite graph. Using $w$, we can project a term-network, $T$, whose edges are modified by this weighting matrix:

$$T = wBB'w', \qquad T_{ij} = \frac{\sum_k B_{ik}B'_{kj}}{\sum_n B_{in} \sum_m B'_{mj}}.$$

In this network the weights of $T$ have a maximum value of one when the same single gene is annotated to both term $i$ and term $j$ and a minimum value of zero when none of the genes annotated to term $i$ are annotated to term $j$. The use of this weighting matrix biases the weights of network edges to those between low degree

terms and therefore the lower branches of the DAG.

## 5.4 Results: An Alternate "Natural" Grouping of GO Terms

### 5.4.1 Comparison with the DAG

Our first goal is to determine if there is any community structure in $T$, meaning are there clusters of terms in $T$ within which there are many or high-weight edges, but between which there are only few or low-weight edges. One way to measure the strength of community structure is through a quantity known as modularity ($Q$) defined as:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j),$$

where $\delta$ is the Kronecker delta function, $c_i$ is the community of node $i$, $k_i$ is the degree of node $i$, $A$ is the adjacency matrix, a matrix with values of representing the weight between nodes $i$ and $j$, and $m$ is the total weight of the edges in the network [60]. This value approaches one for a large amount of community structure, but values greater than about 0.3 are generally accepted as indicative of moderate community structure [61].

To determine the community structure of $T$ we used a weighted version of the Fast Community Structure algorithm [18] implemented in R and determined the communities of terms at maximum modularity. We found a high degree of community structure, with modularity values exceeding 0.55 (Table 5.1). There were also a fairly large number of communities found at these modularity values,

| | Communities Found | Max Modularity | Similarity to GO |
|---|---|---|---|
| Biological Process | 49 | 0.59 | 0.03 |
| Molecular Function | 82 | 0.88 | 0.05 |
| Cellular Component | 39 | 0.80 | 0.05 |
| All Annotations | 53 | 0.60 | 0.03 |

Table 5.1: The community structure properties for $T$ projected in both the combined and individual three ontologies as well as the similarity of the found community structure to a dicretized partitioning of the GO DAG.

indicating that there are indeed many individual modules of terms which share gene annotations.

To quantify the similarity between our found communities and the Gene Ontology hierarchy, we compared the communities of $T$ defined at maximum modularity to branches of the Gene Ontology DAG defined based on second tier terms, by which we mean GO terms whose only parent term is one of the three main ontologies. We compared partitionings using an information theory measurement [24] which varies between 0 and 1:

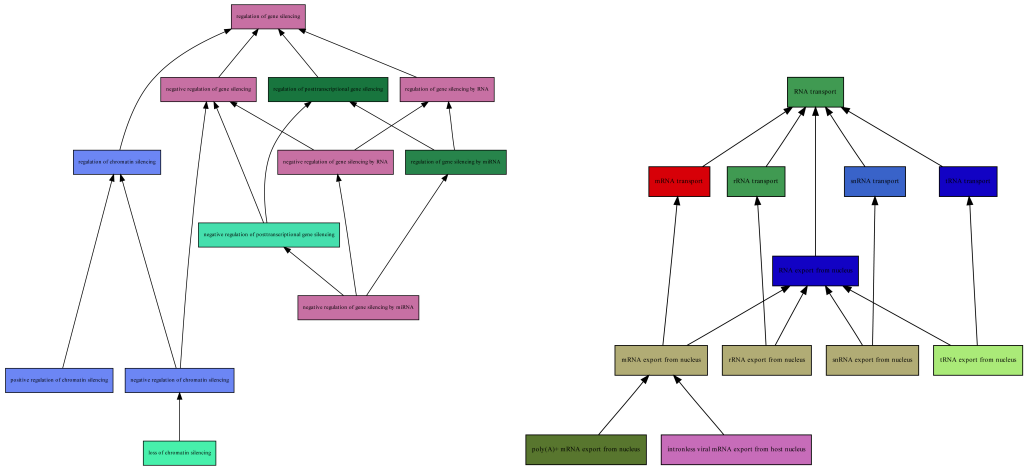
$$I(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}\log\frac{N_{ij}N}{N_{i.}N_{.j}}}{\sum_{i=1}^{c_A}N_{i.}\log\frac{N_{i.}}{N}+\sum_{j=1}^{c_B}N_{j.}\log\frac{N_{.j}}{N}},$$

where $N$ is a confusion matrix, constructed such that each element $N_{ij}$ corresponds to the number of nodes in community $i$ in one partitioning which are in community $j$ in another partitioning. $N_{i.}$ and $N_{.j}$ are the sum over the rows and columns of $N$, respectively. $I(A,B) = 0$ represents completely dissimilar partitionings, and

$I(A, B) = 1$ represents identical partitionings.

Because this measurement requires discreet communities, and branches of the GO hierachy are allowed to be overlapping because of the DAG structure, we developed a discrete term partitioning for the branches in GO, where child terms which belonged to more than one second-tier parent were placed in a community with the parent closest to the child in the DAG structure (see 5.7). Although this is clearly an approximation of the community structure of the DAG, it should be adequate to determine if our found communities are similar or different from the branches of the hierarchy.

With this measure we found that our term communities were very dissimilar from the GO DAG. In order to verify that this dissimilarity was not a consequence of our partitioning of the DAG into discreet communities, we visually illustrated the communities on several branches of the DAG (Figure 5.3). Some similarities can be seen but also many differences between our found communities and the GO DAG. Only a few terms within each branch are identified as belonging to the same community, but the majority of the branches are fractured into many different communities.




(a) GO:0060968 - "regulation of gene silencing"

(b) GO:0050658 - "RNA transport"

Figure 5.3: Communities of terms identified in our projected term network highlighted as a color on the (A) "regulation of gene silencing" and (B) "RNA transport" branches of the GO DAG. Related terms generally belong to the same community as determined by community-detection algorithms. However, even at these lower levels of the DAG, terms within a specific branch of the hierarchy do not necessarily belong to the same community.

### 5.4.2 Using the term communities to evaluate and predict genetic function

#### 5.4.2.1 Evaluating genes associated with term communities

Of course, just because we have found an alternate way to classify gene ontology terms using gene-term annotations it does not imply that this classification will be useful for understanding and interpretting real biological data. However, since the classification schemes relied upon the association of genes with functions, it is reasonable to believe that these communities of terms have biological meaning. To verify whether this alternate classification of terms could be useful in predicting genetic function we calculated the enrichment of gene signatures for cancer in

branches of the gene ontology hierarchy as well as in our communities of terms. For our defined cancer signatures we used the cancer neighborhood and cancer module gene sets published by the Broad Institute [80].

Since a parent term carries all the annotations of its child terms, in order to determine the set of genes to use in evaluating the statistical enrichment of the parent function in another group of genes, all one needs to do is to select the annotations in GO to that parent term and all its children, namely, to identify the set of genes annotated to all the terms within that branch of the hierarchy. Therefore, one might suggest that in order to evaluate our term communities we should first identify the set of genes annotated to all the terms within a community. However, this way of determining gene-associations to our term communities is flawed. For example, any term community which contains one of the top three ontologies will be associated with all the genes which are annotated to terms within that ontology, even if the term community itself is very small and only contains a small subset of terms.

To address this issue, we propose an alternate evaluation method for comparing gene enrichment in communities of terms and branches of the hierarchy which takes advantage of what our communities of terms represent and in the case of the GO hierarchy should approximate the gene annotations currently used. First, imagine how a branch of the DAG would be represented on our network. In the network each branch is a collection of terms connected by edges. Each edge can represent a collection of genes which are common between those two terms. Taking the set of genes which are on any of the edges in this portion of the network results in the set of genes which are annotated to the terms on the branch - excepting the ones

only annotated to the parent term. This is a reasonable exclusion, since the depth of a gene's annotation is related to the confidence of that annotation. In a similar manner we can take the subnetwork of one of our term communities, and identify the set of genes on any of the edges of the subnetwork.

Once we have our sets of genes, we can compare these sets to the set of genes in a gene signature using the same statistical mathematics as is typically applied in many GO analysis software, i.e. hypergeometric probabiliy/Fisher's exact test. For the branches the results should approximate those which are typically produced, except in the case where many annotations to the parent term would normally drive the final results.

#### 5.4.2.2 Enrichment in cancer signatures

After identifying the genes represented in our term communities and the GO branches via the method described, we calculated the probability of overlap between these genes and defined cancer signatures using the hypergeometric probability:

$$p = \sum_{N_v = N'_{12}}^{min[N_1, N_2]} \frac{\binom{N_1}{N_v}\binom{T - N_1}{N_2 - N_v}}{\binom{T}{N_2}},$$

where $p$ is the calculated p-value, $T$ is the total number of genes with GO annotations within the ontology in question, and $N_1$, $N_2$ and $N_{12}$ refer to the number of genes

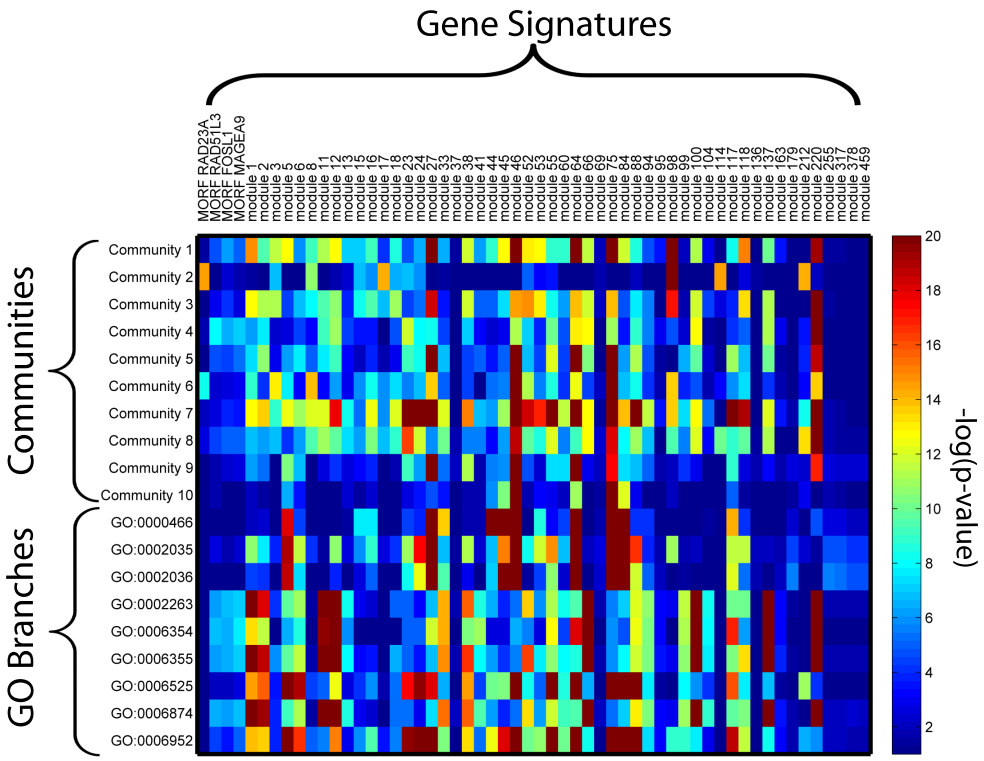



Figure 5.4: A heat map showing the statistical enrichment of cancer signatures in sets of terms. Only signatures with at least 300 members and sets of terms which have statistical correlation with several of these defined signatures are shown. Gene signatures are enriched in both our defined communities of GO terms and branches of the GO DAG. The results of this subset are similar to results for the all signatures and sets of terms investigated.

in a signature, the number of genes in a branch/term community and the number overlapping between those two sets, respectively.

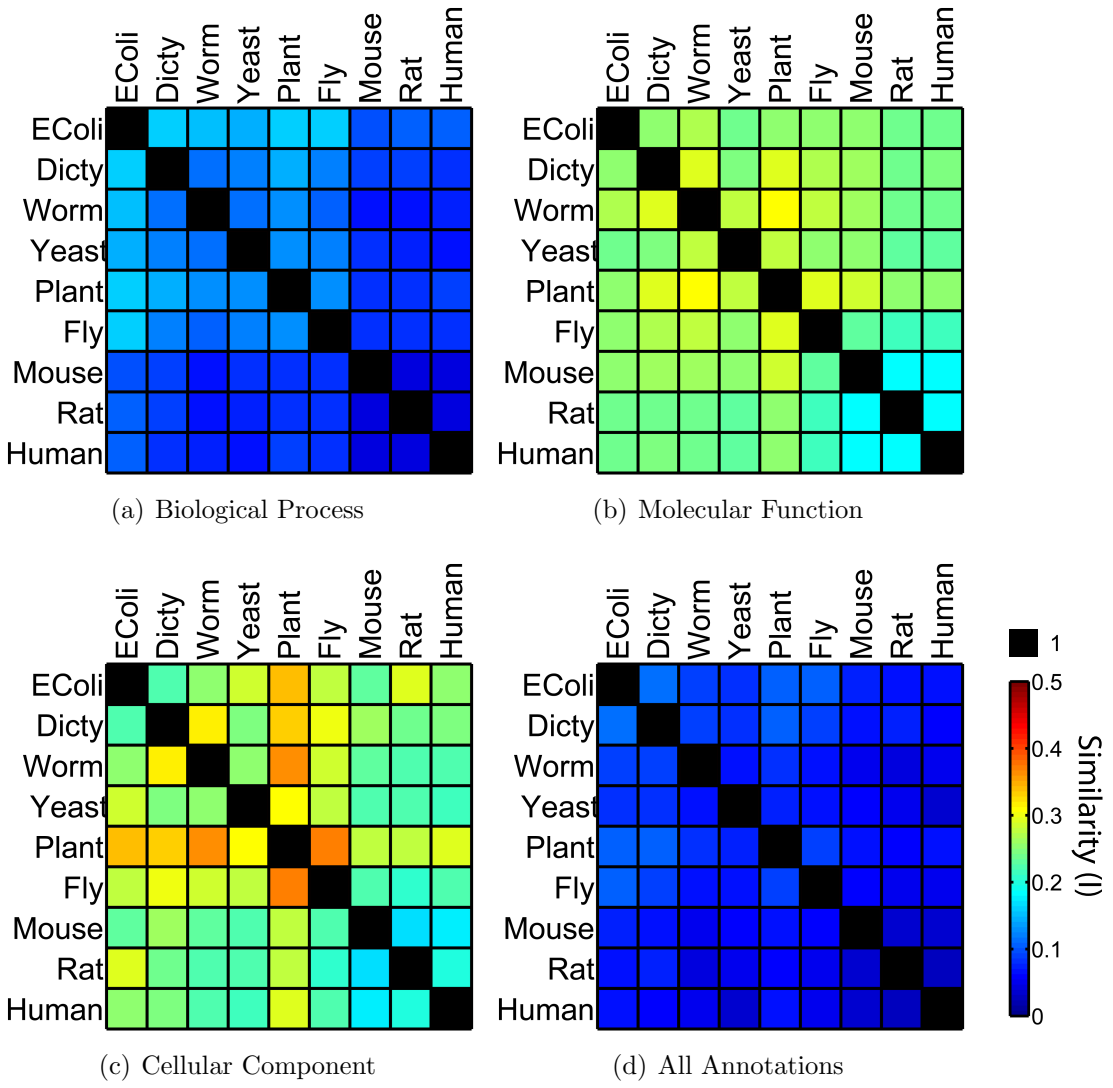
Both GO branches and our communities show a statistical enrichment in cancer signatures, lending biological validity to our term communities. Figure 5.4 shows a heat map of the p-values for a sample of signatures, communities and branches. For simplicity we only chose to show data for signatures with at least 300 members and communites and branches which have a significant p-value in several of these larger signatures.

Both term communities and GO branches are statistically enriched in cancer signatures. On the whole, these cancer signatures show similar levels of enrichment in the term communites and GO branches. The patterns of enrichment are not similar for communities compared to branches, suggesting that these two sets capture different biological information. For example, only a few gene sets are enriched in Community 2, however, these gene sets show at most very mild statistical enrichment in the other communities or GO categories.

### 5.4.3 Species-specific term-networks

Even though the Gene Ontology is designed to establish a species-independent terminology, the communities of terms in the term-networks for various species aren't necessarily the same. Our term networks represent a partitioning of the Gene Ontology in a species-specific manner. In order to determine if our projected term networks are highly species-dependent we compared the partitioning of terms across several species including *E.coli*, Yeast, *Dicty*, *C.elegan*, *Arabidopsis*, *Drosophila*, Mouse, Rat, and Human.

The species-specific term communities are only moderately different from each other with most similarity values varying between 0.1 and 0.3. The "Cellular Component" ontology communities are the most similar across the species. This is most likely a consequence of the relatively small size of the ontology, and the fact that the average degree of the terms in this ontology is high compared to the other two ontologies. These two features combine to create a smaller network with generally




(a) Biological Process

(b) Molecular Function

(c) Cellular Component

(d) All Annotations

Figure 5.5: Similarity between partitions of terms for different species for (A) "Biological Process" annotations, (B) "Molecular Function" annotations, (C) "Cellular Component" annotations, and (D) using all annotations. There is only limited similarity between the paritionings between the different species suggesting that the way biological functions group together in a genetic-framework may be species-dependent.

lower weights between pairs of terms. The "Biological Process" ontology is the most dissimilar between species. It is biologically consistent that the components needed to make a cell should be more similar across species than the organization of biological processes. Of the three ontologies, "Biological Process" also had the highest number of gene annotations, leading to a much denser and more specific organization of terms.

There is also more dissimilarity between mammalian species than between the other organisms analyzed. This is interesting as it could be a consequence of evolutionary diversity playing a more dominant role in the organization of biological function in complex organisms. In constrast, the organization of terms in the only non-animal organism studied, *Arabidopsis*, is more similar to the other species than these species were to each other. This may be due to the fact that this species has more genes listed in the annotation files than the others investigated.

## 5.5 Discussion

The graph structure of the Gene Ontology has never been exploited in a manner that reveals organization of biological function that is unique from the published hierarchical classification. Our method of classifying functional terms produced partitionings which were very different from the Gene Ontology DAG. This implies that our scheme for partitioning the terms represents a natural way in which to form functional modules that rely on genetic behavior. Furthermore, the suggested method takes advantage of a large amount of data from a variety of sources and should

create a classification scheme that is biased only on the type of data reported rather than the organization of human conceptions.

Although our approach creates a natural partitioning of biological function, the biological meaning of the communities of functional terms remains unclear. On a mathematical level these communities should represent sets of very specific biological functions which are generally performed by the same collection of genes. However, labelling and understanding the biological meaning behind these communities is a topic of interest that will require further investigation. As opposed to the DAG, our communities represent a discrete partitioning of the functional terms. However, fuzzy communities could be detected in the term-term network using methods for finding overlapping communities [63]. In addition, hierarchical structure could be identified from the term-term network [19][20]. We leave these two directions for future work.

## 5.6 Conclusion

By using functional annotation data within the Gene Ontology we were able to construct an alternate, natural, and biologically-relevant way in which to categorize cellular functions. This categorization is structurally and conceptually distinct from the GO DAG and allows for functional relationships between terms which do not share a parent/child relationship. Since we built our communities based on the specificity of the information shared between terms, they have the potential to be incredibly useful in understanding the collective behavoir of genes, especially at a

species-specific level.

## 5.7 Notes

### 5.7.1 Using annotation files to construct the bipartite graph

The annotation files used in this analysis are provided by the Gene Ontology Consortium and are freely available from the Gene Ontology website (www.geneontology.org). These files only list "lowest order" term-gene associations, meaning that if a gene and term are associated, that gene is also presumed to be associated with all the parent-terms of the listed term. If two annotations along the same branch appear, it is assumed that the "higher" annotation is correct whereas there is some uncertainty in the lower annotation. In constructing our bipartite graph we took the "higher" or more confident annotations listed and used them to constuct the set of all possible gene-term annotations, propagating up the DAG along both "is-a" and "part-of" relationships.

### 5.7.2 Partitioning the DAG

When comparing the partition of our term-term networks to the GO DAG we began by separating the DAG into its the three main ontologies: "Biological Process," "Cellular Component" and "Molecular Function." When looking at the next lower level partition of the GO DAG we partitioned the DAG into groups based on those terms whose only parent term was one of the three main ontologies. Since not all terms in the GO annotation uniquely belong to one of these second-

tier terms, we developed a discrete partitioning method in which we calculated the shortest paths, or the geodesics, between each child term and all 2nd-tier parents terms. Children were put into a community with the parent with whom they had a shorest geodesic. If the child had the same geodesic to more than one parent, the assigned community was randomly chosen from that set of parents.

# Appendix A

# Background Biology and Experimental Methods

## A.1 The Genome

Every living organism on this planet, from the simplest bacterium to the most advanced human being, contains the primary instructions for its existence in a four letter code. This four letter code is found in the deoxyribonucleic acid, or the DNA, contained in the organism's cells. In most organisms, the four components that make up DNA are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). These four "bases" link together via hydrogren bonds to form "base-pairs" (bp) with Adenine always across from Thymine, and Cytosine always across from Guanine. These base-pairs connect to each other along a phosphate backbone, resulting in a long molecular chain encoding biological information. Because of its biochemical properties, this chain tends to twist into a double-helix structure. This twisting has an approximate period of ten base-pairs (3.3 nm) and produces two "grooves." The major groove is the larger of the two and is approximately 2.2 nm wide. The minor groove is approximately 1.2 nm wide.

The Genome refers to all the genes in an organism. Mammalian genomes have between twenty and twenty-five thousand genes. A gene is a region of DNA which contains the instructions to make a protein. These proteins perform many diverse duties. Some proteins are responsible for activities within the cell, and others are

involved in continued gene regulation. Some proteins are molecules used to turn genes on and off. For example, RNA Polymerase II (RNAP) is the primary protein responsible for transcription, the process by which genetic information is extracted for use elsewhere in the cell.

## A.2 Regulation of Gene Expression

Gene regulation refers to the mechanisms through which gene expression levels are controlled. Recent developments have strongly suggested that regulation is a complex dynamical process with highly integrated feedback and control mechanisms that result in diverse expression patterns [25].

### A.2.1 Promoters and regulation by transcription factors

Many genes are known to be at least in part regulated by regions of DNA sequence called promoters, located upstream of the transcriptional start site (TSS). The TSS is the location along the DNA where transcription of a gene begins. The full promoter region of a gene can extend upwards of 1000 bp before the TSS of a gene, however, the proximal promoter, extending only 200 bp before the TSS, typically contains the majority of DNA binding sites necessary for regulatory proteins such as transcription factors (TF), to bind to the DNA and initiate (or prevent) transcription of that gene.

One common class of proteins is that of basic leucine zipper (bZip) transcription factors. Examples include CREB and C/EBP. bZip transcription factors con-

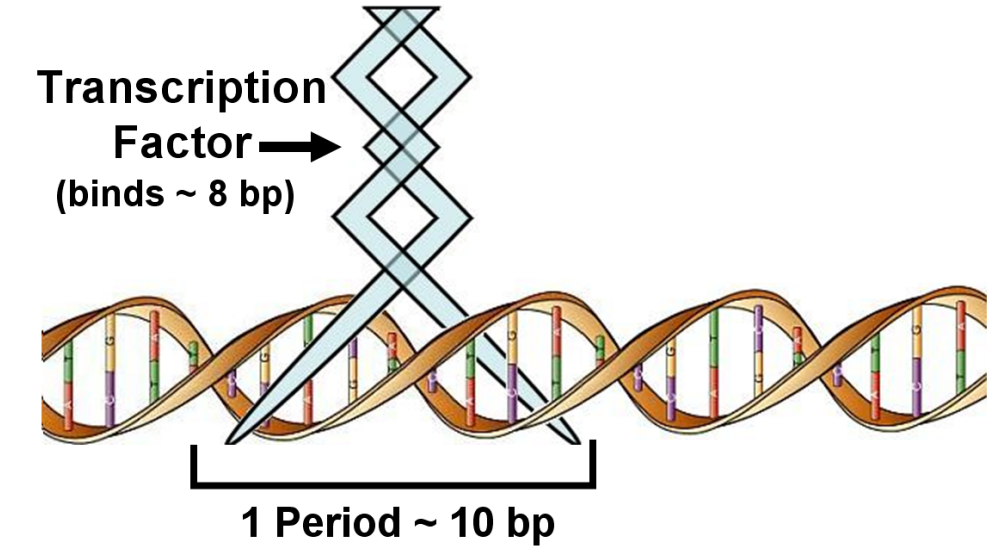



Figure A.1: An illustration of how the basic DNA-binding region of a bZip transcription factor protein binds into the major groove of DNA.

tain two $\alpha$-helixes (coiled structures formed of amino acids) which dimerize, or join together, via a hydrophobic leucine zipper, and a basic DNA-binding region which interacts with the major groove of DNA through hydrogen bonding (Figure A.1).[1] If the two $\alpha$-helixes in the bZip are the same, the bZip is a homodimer; if they are different, the bZip is a heterodimer. Homodimers are unique in that they generally bind palindromic regions of DNA, with each half of the palidrome interacting with one of the two $\alpha$-helixes. This is the case with CREB and C/EBP whose canonical binding sites are the palindromes TGACGTCA and TTCCGGAA respectively.

Although they are physically much larger, on the order of tens of nano-meters, TFs typically bind to DNA sites that are between 5 and 10 bp in length (2-3 nm). By binding both to each other and to the DNA, transcription factors form what are

[1]Image adapted from "The Science Creative Quarterly." (www.scq.ubc.ca/decoding_icelands_dna/)

known as transcriptional complexes over the core promoter region of a gene. These complexes generally stretch from around -50 bp to +20bp around the TSS and may recruit or include RNA Polymerase II, the enzyme which catalyzes the transcription of DNA.

### A.2.2 Epigenetic regulation

Epigenetics refers to the regulation of gene expression by mechanisms other than the DNA sequence. Epigenetic mechanisms include histone modifications, DNA methylation, and nucleosome positioning.

In mammalian genomes, such as human or mice, the DNA in a single cell is about three billion base-pairs long, which if laid out straight would be approximately one meter in length. However, DNA in the cell is normally found tightly coiled into structures known as chromatin. Approximately 147 bp of DNA wraps 1.67 times around groups of eight histones, a type of protein, to form nucleosomes [53]. A segment of 20-60 bp of "linker" DNA joins these nucleosomes together it what looks like "beads on a string," which can then pack closely together, forming a tight structure.

Because every cell in an organism shares the same DNA, or base-pair genetic code, this structure is one way by which gene expression can be controlled differently in individual tissues. Nucleosome occupancy serves as a formidable barrier to the binding of TFs and other regulatory proteins to DNA and represses gene expression [46]. However, the expression of genes can be controlled on an epigenetic

level by histone modifications and DNA methylation. When a histone undergoes modification, the way DNA is wrapped around that histone changes, facilitating or impeding gene translation and therefore gene expression.

### A.2.3 CpG Islands

When occurring adjacent to a Guanine, the Cytosine nucleotide in genomic DNA can be covalently methylated by adding a methyl group. This modification lies in the major groove of DNA. Methylation of the CpG (Cytosine-phosphate-Guanine) is generally thought to be associated with gene repression and it has been shown that aberrant methylation plays a significant role in a number of diseases including cancer.

In mammalian genomes, the CpG dinucleotide is rare, occuring at only 20% of the expected frequency [81] and is typically methylated [10][9]. The exception is in CpG Islands (CGI). CpG Islands are defined as regions in the DNA at least 200 bp long where C+G comprise more than 50% of the nucleotides and CpG dinucleotides occur at greater than 60% the expected frequency [32]:

$$f_C + f_G > 0.5$$

$$f_{CG}/(f_C * f_G) > 0.6$$

CpG Islands often occur in regulatory regions of the DNA such as promoters and their occurrence is known to correlate with the recruitment of RNAP and gene

activation [43], especially in constitutively active "housekeeping" genes which are bound by RNAP in multiple tissues.

## A.3 Methods for Interogating Biological Systems: Chromatin Immunoprecipitation (ChIP-chip)

Chromatin Immunoprecipitation microarray assays (ChIP-chip) make it possible to screen for binding of a protein to promoter sites on a genome-wide basis. The procedure is *in vivo* since it interrogates the properties of living cells. An illustrated overview of the process is provided in Figure A.2.[2]

### A.3.1 Chromatin-Immunoprecipitation

Chromatin immunoprecipitation (ChIP) is an experimental procedure used in order to determine the DNA bound by a protein on a genome-wide level. The process begins by crosslinking, or taking live cells and adding formaldehyde in order to attach the proteins currently in the cell to the nearest DNA. In the majority of cases, this affixes a protein to the DNA to which it is bound. DNA is then extracted from the sample and it is sonicated to break it into fragments on the order of 200-1000 bp. Some of these fragments will have protein bound to them and others will not. Once the DNA is broken into fragments the sample is separated into two, half of which will be used as a "signal" to evaluate the DNA bound by a particular protein and the other half as a "background" control sample of the DNA in the cell

[2]Image created using cells modified from http://www.cellapplications.com/animal-cells-rat-cells.php and petri dish from http://pgrsource.com/pgr_store/agora.cgi?6482349.15011&product=Labware

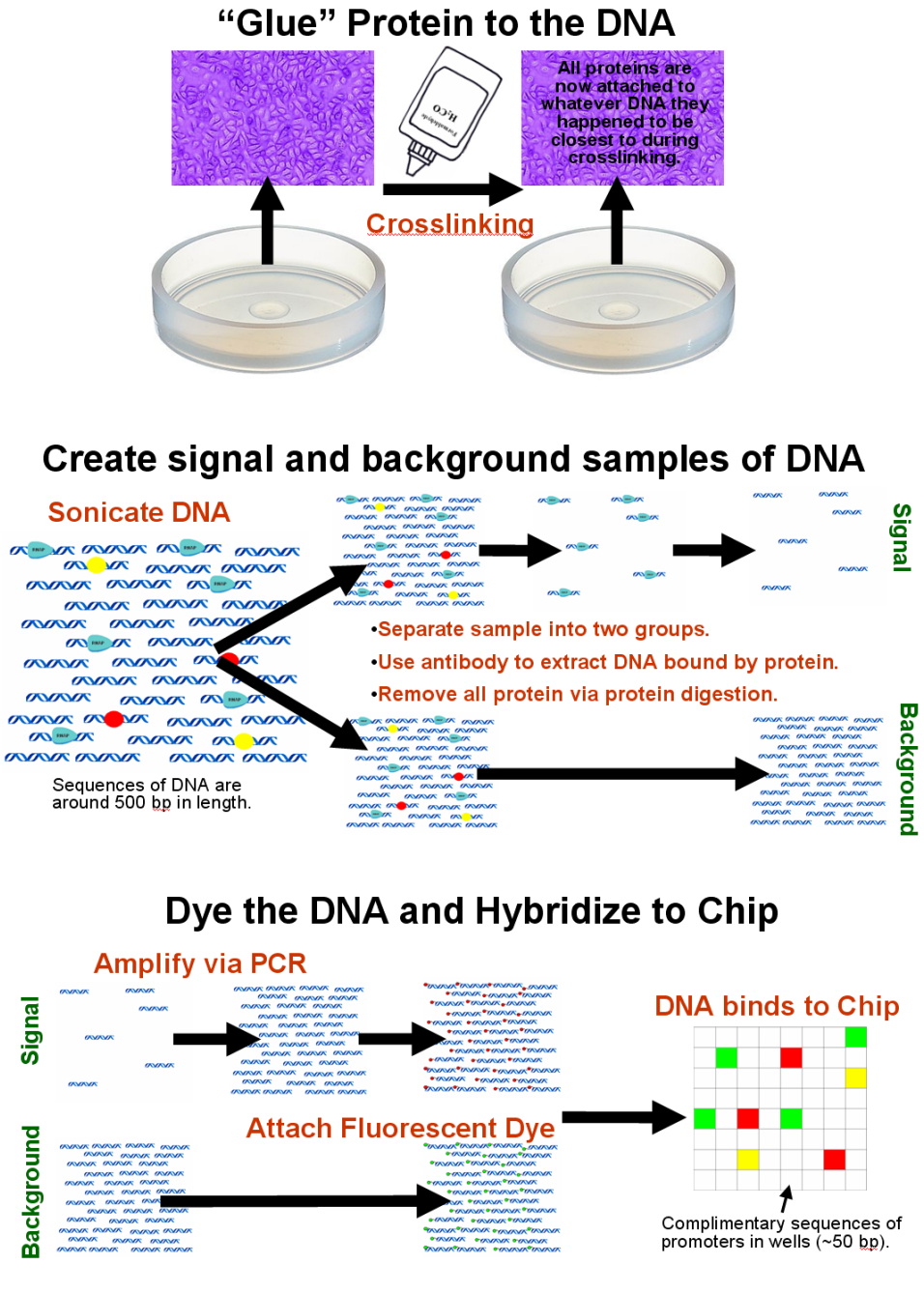



Figure A.2: A visual summary of the ChIP-chip protocol.

population.

Antibodies are gobulin proteins which recognize and bind to specific proteins in the cell. Antibodies attached to beads which specifically recognize the protein of interest are added to the "signal" sample. The DNA sequences which are bound by the protein of interest are then extracted by selecting the DNA attached to the beads in the sample. Protease is then used to digest the protein in both the signal and background samples, leaving only the DNA. Once the protein is removed from the samples, the signal sample is amplified, or enlarged, via PCR to approximately the same size as the background sample.

### A.3.2 Microarray technology and analysis

For ChIP-chip, the chromatin immunoprecipitation procedure is followed by microarray, or "chip" technology. Both the "signal" and "control" samples are denatured to create single-stranded DNA and different colored fluorescent dyes are attached to the DNA in the two samples. Each sample is then added to a microarray which contains complimentary single-stranded DNA. If DNA from the samples matches that on the microarray, then the sample DNA will bind at that location on the microarray. Once this process is complete, the microarray is scanned to determine the relative amount of DNA from the two samples bound at each spot on the microarray. This information is then mapped back to the genome.

The microarray is typically laid out in a grid format, with each square or "well" containing a dense collection of the same single-stranded segment of DNA,

called an oligo or probe, on the order of 50-100 bp in length. These segments of DNA are a selection from the regions of the genome which are under investigation, for example, the promoter regions of genes. Current microarrays can contain up to two million probes allowing for a fairly dense coverage of selected regions of the genome.

# Appendix B

# Consequences of CpG methylation of CRE-like sequences

## B.1 Overview

Methylation of the CpG dinucleotide in promoters is typically associated with gene inactivation because it both recruits repressor proteins and inhibits the DNA binding of activating transcription factors. However, recent data indicate that some methylated promoters are transcriptionally active. A ChIP-chip promoter analysis of primary new-born keratinocytes from mouse skin indicates that C/EBP$\alpha$ binds methylated promoters containing CRE-like sequences. These results identify a new role for the methylated CpG, creation of a transcription factor binding site (TFBS) that is critical for initiation of tissue specific gene activation. This change in TF binding specificity with methylation suggests a switch with C/EBP$\alpha$ initiating activation of tissue specific methylated promoters and their subsequent demethylation inducing CREB binding and maintenance of gene activity in long lived differentiated cells.

## B.2 Background

In vertebrate genomes, the CpG dinucleotide is rare and occurs most often in the promoters of constitutively active housekeeping genes where it is found in

clusters called CpG Islands [11][32][82]. In the early embryo, the CpG dinucleotide is unmethylated. During the blastula stage the cytosine becomes methylated in non-housekeeping genes but the promoters of housekeeping genes remain unmethylated [9]. Promoters with methylated CpGs are typically silent and are bound by methyl binding proteins (MBP) which have been implicated in promoter repression [12]. Furthermore, CpG methylation prevents DNA binding of many transcription factors (TFs) including SP1 [71], CREB [88], ETS [71][33], NRF-1 [71][17], Myc|Max [65], and AP2 [21]. These TFs typically bind housekeeping promoters suggesting that in housekeeping genes unmethylated CpGs are critical for TF binding [71] and aberrant CpG methylation of housekeeping genes, as occurs in many cancers [42], could inhibit transcription by directly inhibiting TF binding. There are reports of enhancers of tissue specific genes that become demethylated when transcriptionally active [13] lending support to the suggestion that CpG methylation is a general repressive epigenetic mark in vertebrate genomes [12].

The suggestion that CpG methylation and transcription are mutually exclusive has been cast in doubt with recent global analyses that have identified promoters that are both methylated and transcriptionally active [26][87]. Furthermore, one study showed that the CpG methylation pattern is not as dynamic as previously thought [26] and another revealed that it is mainly regions outside of promoters that become demethylated upon cellular differentiation [58]. For example, only 20% of lung specific transcripts have a 5 ' UTR that becomes demethylated upon tissue specific expression [22], implying that some promoters are likely methylated and active.

## B.3 Results

### B.3.1 Promoters fall into two distinct classes

These results led us to investigate whether instead of being only a transcriptionally repressive mark, CpG methylation might facilitate transcription factor (TF) binding to DNA in certain instances. We therefore investigated the effect of CpG methylation on transcription factor binding in differentiating primary new-born mouse keratinocytes. The CpG methylation status of promoters, including tissue specific promoters, was determined using methylated DNA immunoprecipitation (MeDIP) [86]. Plotting the amount of methyl CpG at a promoter versus the number of CpGs in the promoter region (-1,000 bps to +500 bps) produces two groups of promoters (Figure B.1 A). One group (∼9,000 members) has fewer CpGs with a steady increase in methylation as the number of CpGs increases; presumably these promoters are completely methylated [26]. The second group of promoters of similar size (∼10,000 members) has a larger number of CpGs but fewer methyl CpGs suggesting they are either partially or completely unmethylated. These results are consistent with what is observed in human primary fibroblasts [87]. A heat map of RNAP binding in undifferentiated keratinocytes indicates that RNAP binds primarily unmethylated promoters; in the top 40% of promoters bound by RNAP, over 95% are unmethylated (Figure B.1 B).

We determined the methylation status of each promoter based on whether or not that promoter appeared in the top diagonal cloud (methylated) or the bottom circular cloud (unmethylated) in the Number of CpGs vs. MeDIP graph. The line

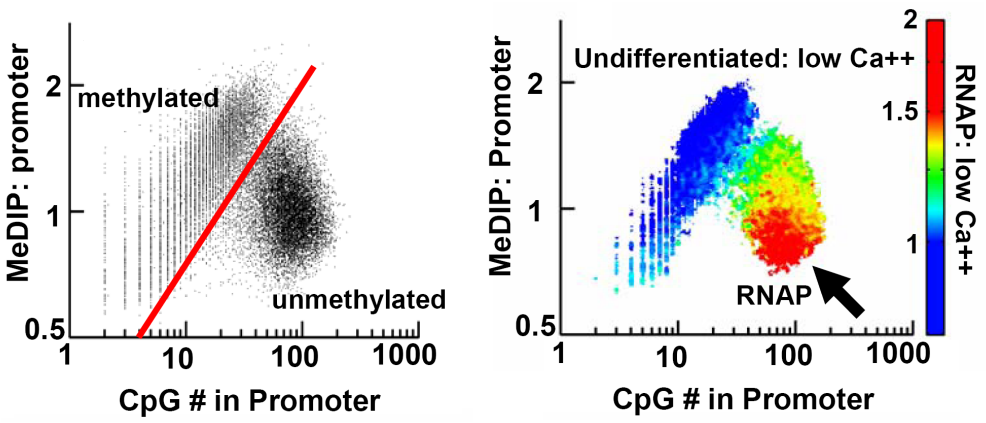



Figure B.1: Plotting MeDIP values vs. the number of CpGs in promoters reveals two distinct classes of promoters.(A) Methyl-CpG dinucleotide abundance in undifferentiated keratinocytes vs. the number of CpG dinucleotides in a promoter (-1,000 bps and +500 bps). Two distinct groups of promoters are distinguishable. A line drawn between these groups identifies 8,863 methylated promoters and 11,465 unmethylated and partially methylated promoters. (B) Heat map of RNAP binding in undifferentiated keratinocytes on plot of MeDIP vs. CpG dinucleotide number in a promoter.

dividing the two clouds was determined as follows: First, a line was fit through the top cloud based on where the density in that cloud peaked. This line was then used to transform the methylation data through a rotation:

$$x' = \frac{mx - y}{\sqrt{1 + m^2}},$$

where $x'$ is the transformed methylation data, $(x, y)$ are the original coordinates of the (methylation, Number of CpG) point, and $m$ is the slope of the fit line. The transformed data was plotted as a histogram and Gaussians were fit to each of the two resulting peaks. The intersection of these two Gaussians was calculated. Promoters with transformed values above the intersection were considered

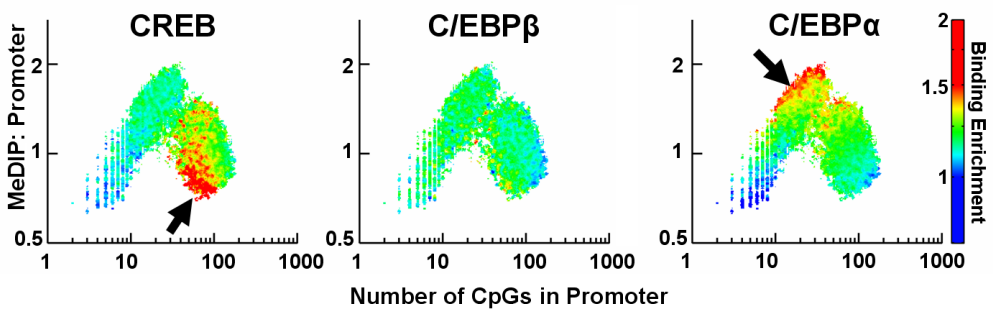



Figure B.2: Localization of B) CREB, C) C/EBP$\beta$, and D) C/EBP$\alpha$ to promoters in primary keratinocytes plotted as a heat map on methyl-CpG promoter content determined by MeDIP vs. CpG dinucleotide number in the promoter (-1,000 bp to +500 bp). CREB preferentially binds to promoters that are unmethylated, while C/EBP preferentially binds to promoters that are methylated.

non-methylated and those below were considered methylated. For keratinocytes, this method identified 8,863 methylated promoters.

### B.3.2 Effect of methylation on TF binding

To determine whether a methylation dependent difference in DNA binding specificity occurs in vivo, we used ChIP-chip experiments in keratinocytes to determine the promoter localization for CREB, C/EBP$\beta$, and C/EBP$\alpha$ in the genome. CREB binds unmethylated promoters, C/EBP$\beta$ binds both unmethylated and methylated promoters while C/EBP$\alpha$ binds to similar promoters as C/EBP$\beta$ but also binds some additional methylated promoters (Figure B.2).

Although this shows in a global way that C/EBP$\alpha$ binds to methylated promoters and that RNAP and CREB binds to unmethylated promoters, we wish to address in more detail which individual sequences each of these transcription factors binds. In order to do this we calculated the 8mer-association-with-C/EBP$\alpha$

and 8mer-association-with-CREB.

For a particular 8-mer, this quantity ($b_8$) is the average binding of the transcription factor to promoters that contain that 8-mer, normalized by the average binding to all common promoters ($\bar{b_p}$).

$$b_8 = \frac{\sum_p b_p \delta_{8p}}{\bar{b_p} \sum_p \delta_{8p}},$$

where $p$ is the promoter in question. $\delta_{8p}$ is equal to one if the 8-mer occurs in the promoter sequence and zero otherwise. To obtain distinct values for methylated and unmethylated CpG-containing 8mers, all the CpGs in promoters in the methylated group were considered methylated and therefore distinct from the CpGs in promoters in the unmethylated group. This was done by transforming all the CpGs in the methylated promoters to mCpGs before doing the association calculation.

CREB, as expected, is generally associated with unmethylated 8mers, while C/EBP$\alpha$ is generally associated with methylated 8mers (Figure B.3). The canonical binding sites for CREB (TGACGTCA) and C/EBP$\alpha$ are highly enriched in their respective TF experiment. One interesting feature is that only the unmethylated version of the canonical CREB binding site (TGACGTCA) is enriched in promoters bound by CREB, whereas both the unmethylated and unmethylated versions of the canonical C/EBP$\alpha$ binding site (TTGCGCAA) are enriched in promoters bound by C/EBP$\alpha$. This suggests that the C/EBP$\alpha$ protein can bind to DNA irrespective of the methylation status of the DNA.

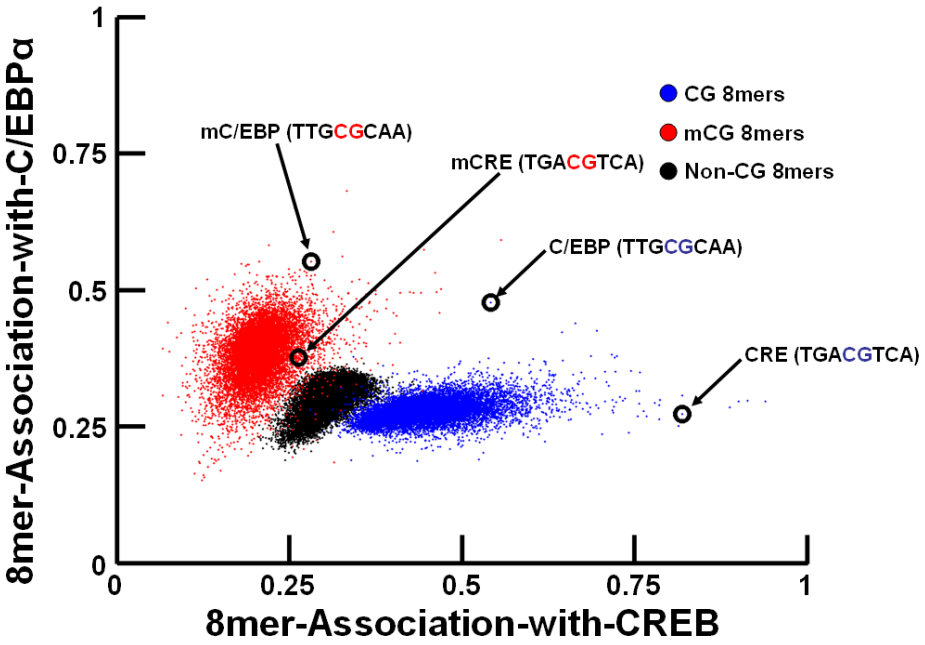



Figure B.3: A plot of the 8mer-Association-with-C/EBP$\alpha$ versus the 8mer-Association-with-CREB showing methylated and unmethylated sequences separately. The 8mers most associated with C/EBP$\alpha$ generally contain a methylated-CpG (mCpG) while the 8mers most associated with CREB generally contain unmethylated CpGs. Also marked are the canonical binding sites for CREB (TGACGTCA) and C/EBP$\alpha$ (TTGCGCAA). Only the unmethylated CRE-binding site is enriched in promoters bound by CREB, however, both the methylated and unmethylated versions of the C/EBP$\alpha$ binding site are enriched in promoters bound by C/EBP$\alpha$.

This change in TF binding specificity with methylation suggests a switch with C/EBP$\alpha$ initiating activation of tissue specific methylated promoters as is observed in short term keratinocyte cultures and their subsequent demethylation inducing CREB binding and maintenance of gene activity in long lived differentiated cells as is observed in both liver and heart tissue.

## B.4 Conclusion

In summary, we show that a methylated CpG dinucleotide in DNA sequences can act as a TFBS for the C/EBP$\alpha$ transcription factor. This potential for methylation of CpG to create a TFBS for TFs that activate gene expression augments its more familiar role in the repression of gene expression. Consistent with the idea that CpG methylation creates a TFBS for C/EBP$\alpha$ that is essential for the activation of some tissue specific genes, demethylation of the genome facilitates the conversion of fibroblasts into stem cells that are unable to differentiate [58].

## B.5 Methods

### B.5.1 Determining the Promoter Set

We limited our promoter analysis to those 24,844 promoters not located on the X or Y chromosome. 3,940 promoters were further eliminated that either had unknown DNA segments larger than 150 bp or which contained large regions of DNA sequence ($\geq$150 bp) which were identical to regions in 10 or more other promoters. For the remaining set, if several promoters all shared large regions of DNA sequence

in common (more than 7 unique 150 bp segments) we averaged their binding values and assigned this average to the earliest appearing promoter, leaving us with a final promoter set of 20,328 promoters.

### B.5.2 3-dimensional representations of biological data

Binding heat maps were created by identifying promoters within an elliptical region around a point on a graph and coloring that point based on the resulting median binding of those promoters. The center of each elliptical region was determined by incremental steps in both the X and Y directions. Data was only calculated and plotted for elliptical regions which contained ten or more promoters.

# Appendix C

## Ways to Weight the Projected Term Network

In addition to the weighting scheme proposed in Chapter 5, there are several alternate ways to weight our projected term network ($T$) which one might believe would better recapitulate the structure of the Gene Ontology DAG. Two of these ways are outlined in this section along with a brief discussion of their similarity to the GO DAG.

## C.1 Two Alternate Ways to Weight $T$

### C.1.1 Weighting by shared annotations

From our bipartite graph ($B$) we could immediately generate a projected network relating the ontology terms ($T^{shrd}$):

$$T^{shrd} = BB', \qquad T^{shrd}_{ij} = \sum_k B_{ik} B'_{kj}.$$

Term-term connections in this projected network would have a weight equal to the number of shared genes between the two terms. This method is analogous to the one outlined in [39]. However, as previously discussed, some terms have many gene annotations. Under this scheme two terms that both have many genes annotated to them could easily have a high weight even if that overlap is what one might expect

by chance.

### C.1.2 Normalized by minimum degree

One could also consider weighting each term-term connection by the maximum number of gene annotations those two terms could share, or the minimum degree of the two terms. In that case the values of $T_{ij}$ would take on the values:

$$T_{ij}^{min} = \frac{\sum_k B_{ik} B'_{kj}}{min[\sum_n B_{in} \sum_m B'_{mj}]}.$$

In this scenario the values of $T_{ij}$ would take on a maximal value of one when all of the annotations to one term are shared with the other term, and a minimal value of zero when no gene annotations are shared between the two terms. This means that every child term will have a maximally weighted edge between it and all of its parent terms and vice versus. This alternate weighting *should* bias the resultant network to form modules similar to the branches of the hierarchy.

### C.1.3 Normalized by the product of degree

For reference, in chapter 5 we chose to compensate for the skewed term degree distribution by introducing a diagonal weighting matrix, with off-diagonal elements

equal to zero and diagonal elements equal to:

$$w_{jj} = \frac{1}{\sum_{i=1}^{n_T} B_{ij}},$$

or simply one over the degree of the term in the bipartite graph. $T$ was then:

$$T^{prod} = wBB'w', \qquad T_{ij}^{prod} = \frac{\sum_k B_{ik}B'_{kj}}{\sum_n B_{in} \sum_m B'_{mj}}.$$

This is equivalent to dividing the number of gene annotations shared between term $i$ and $j$ by the product of their degree.

## C.2 Consequences of weighting

Because gene annotations in GO propagate up the hierarchy, when there is a link between two terms in the hierarchy, there is by default a link between those terms in our projected network. In fact, in our projected networks there is always a link between a term and every one of its parent and child terms. Because of this, one might suspect that terms should always fall into communities very similar to the defined DAG structure. Even though the only parameter we vary in our three weighting schemes is how much significance should be given to each edge, this can actually play a very large role in the final discovered term communities.

There are several issues which could lead to the dissimilarity between communities of terms in $T$ and branches of the hierarchy. Firstly, when a gene is annotated

to two terms in separate branches of the hierarchy, then those two terms not only gain a link between them, but a link is added between all the parents of those two terms, leading to many connections between terms in disparate parts of the hierarchy. The effect these links play in grouping the terms into communities can be controlled by how we choose to weight our links.

An initial suggestion may be to weight links between terms such that if two terms share many genes in common, as is expected if one is the parent/child of the other, then that link gains more weight. This is our "shared" weighting scheme. However, the number of shared genes between two terms is not the best indicator that those terms are on the same branch of the hierarchy. For example, say term A, with 3 annotations, is the child of term B with 10 annotations. The number of genes shared between term A and B is, due to the nature of the hierarchy, 3. However, if term C, from a different branch of the hierarchy has 100 annotations, 6 of which are shared with B, then the link between B and C will have a weight of 6 and be twice as strong as the link between A and B, potentially leading to B and C being placed in the same community rather than A and B.

However, we can modify our weighting to take into account the degree of terms. We would like two terms which share only a few annotations but that each only have few annotations to have a stronger connection than two terms which share many annotations but that also each have many annotations. Weighting by the minimum degree will force all weightings between terms within the same branch of the DAG to be equal to one. Cross-branch edges can vary between 0, when there are no shared annotations, to 1, when all annotations from the minimum degree

term are shared with the maximal degree term. As an example, take again terms A, B and C, where A, with 3 annotations, is a child of B, with 10 annotations, and C, with 100 annotations, is from a different branch of the hierarchy. Under our "minimum degree" weighting scheme, because A is a child of B, they share the maximum number of possible annotations in common, and thus the weight of their link in our term network would be one, namely, it would be the number of shared annotations (3), divided by the maximum possible shared annotations, $min([10, 3]) = 3$. On the other hand, the link between B and C would have a weight of only 0.6, or, the number of shared annotations (6), divided by the maximum possible, 10. This weighting scheme should bias our found communities toward branches of the hierarchy, independent of the number of annotations within those branches.

Chapter 5 described a weighting scheme which weighted links by the product of the two term degrees. The product weighting scheme should bias edge weights toward edges between terms in the lower branches of the DAG, or those terms with few gene annotations. If a term has a high degree, then we may expect many shared annotations between that term and other terms in the DAG due to chance. This weighting scheme differs from the above by purposefully given weights between a term with a high degree and other terms a lower weighting such that even if two terms share the maximum possible number of annotations, if one of those two terms has a high degree, the link between the two may be relatively low. Links between two terms will at most have a weight of one over the maximum degree of the two terms. This value can vary from a maximum value of one, when two terms each have

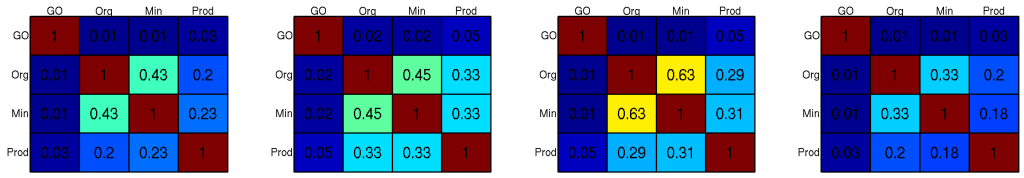
| | Shared Annotations | Normalized by Minimum | Normalized by Product |
|---|---|---|---|
| Biological Process | 0.185793 (3) | 0.2009138 (6) | 0.5864669 (49) |
| Molecular Function | 0.3053261 (7) | 0.3981552 (6) | 0.8797853 (82) |
| Cellular Component | 0.1459397 (3) | 0.2422619 (3) | 0.8029135 (39) |
| All Annotations | 0.1720733 (3) | 0.2070969 (11) | 0.5997856 (53) |

Table C.1: Maximum modularity of network as determined by the Fast Greedy Community Structure detection algorithm as well as the number of communities found at that modularity. Modularity can vary between 0 and 1. The product of degree weighting scheme creates a much more modular network than the other two schemes as well as many more communities of terms.

exactly the same single annotation, to zero, when the two terms share no common annotations.

## C.3 Comparison of Weighting Schemes to the DAG

We determined the community structure at maximum modularity for each weighting scheme using the fast-greedy community structure algorithm [18]. For the individual ontologies, in the case of weighting by annotations and normalizing by minimum degree, there is only a little community structure apparent in $T$. The maximum modularity values fall between approximately 0.1 and 0.4 and only a few communities are found at these modularity values. However, when we weight by the product of the term degrees, there is much more community structure. The maximum modularity values are much higher as are the number of found communities (see Table C.1). Results are similar if we take all gene-term annotations to determine $T$ instead of only those within a single ontology.




(a) Biological Process (b) Molecular Function (c) Cellular Component (d) All Annotations

Figure C.1: Similarity between partitions using GO and different weighting schemes for (A) "Biological Process" annotations, (B) "Molecular Function" annotations, (C) "Cellular Component" annotations, and (D) using all annotations. Although there is some similarity between the three weighting schemes, and especially between weighting by shared annotations (org) and normalized my minimum degree (min), there is little similarity between the communities found under any weighting scheme and the gene ontology branches.

None of our three potential weighting schemes produce communities of terms which are similar to the branches of the GO hierarchy (Figure C.1). Although when weighting by the degree of terms, the similarity to the GO hierarchy is slightly higher, the value is so low it is difficult to tell whether this slight improvement over the two weighting schemes is believable. Even though the shared and minimum weighting scheme found far fewer communities, these communities are still very different from the branches of the DAG (Figure C.2). On the other hand, when compared to each other, the three weighting schemes produce communities which, although not very similar, still have some degree of similarity. This suggests that the structure within $T$ is fairly robust and evident even when the strength of the connections in $T$ is altered to try to favor the GO hierarchy.

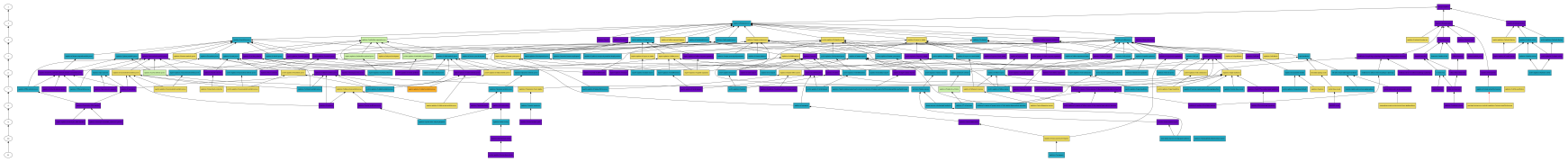


Figure C.2: Communities of terms found by running the fast greedy community-detection algorithm on our projected term network normalized by the minimum degree highlighted as a color on the "biological regulation" branch of the gene ontology hierarchy. Notice how different the found communities are from the GO DAG even for this weighting scheme which should contain many between branch edges of maximal weight.

# Bibliography


[1] R. Albert. Scale-free networks in cell biology. *J Cell Sci*, 118(Pt 21):4947–57+, 2005.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature genetics*, 25(1):25–29, May 2000.

[3] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36+, 1994.

[4] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[5] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, 2007.

[6] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human b cells. *Nat Genet*, 37(4):382–90+, 2005.

[7] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, 3rd Estep, P. W., and M. L. Bulyk. Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–35, 2006.

[8] M. Bina, P. Wyss, W. Ren, W. Szpankowski, E. Thomas, R. Randhawa, S. Reddy, P. M. John, E. I. Pares-Matos, A. Stein, H. Xu, and S. A. Lazarus. Exploring the characteristics of sequence elements in proximal promoters of human genes. *Genomics*, 84(6):929–40, 2004.

[9] A. Bird. Dna methylation patterns and epigenetic memory. *Genes Dev*, 16(1):6–21, 2002.

[10] A. Bird, M. Taggart, M. Frommer, O. J. Miller, and D. Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell*, 40(1):91–9, 1985.

[11] A. P. Bird. Cpg-rich islands and the function of dna methylation. *Nature*, 321(6067):209–13, 1986.

[12] A. P. Bird and A. P. Wolffe. Methylation-induced repression–belts, braces, and chromatin. *Cell*, 99(5):451–4, 1999.

[13] Brian P. Brunk, David J. Goldhamer, Charles P. Emerson, and Jr. Regulated demethylation of themyoddistal enhancer during skeletal myogenesis. *Developmental Biology*, 177(2):490 – 503, 1996.

[14] M. L. Bulyk, A. M. McGuire, N. Masuda, and G. M. Church. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in escherichia coli. *Genome Res*, 14(2):201–8+, 2004.

[15] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, pages 418–29+, 2000.

[16] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–35, 2006.

[17] Y. S. Choi, S. Kim, H. Kyu Lee, K. U. Lee, and Y. K. Pak. In vitro methylation of nuclear respiratory factor-1 binding site suppresses the promoter activity of mitochondrial transcription factor a. *Biochem Biophys Res Commun*, 314(1):118–22, 2004.

[18] A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(6 Pt 2):066111+, 2004.

[19] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Structural inference of hierarchies in networks, Oct 2006.

[20] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, May 2008.

[21] M. Comb and H. M. Goodman. Cpg methylation inhibits proenkephalin gene expression and binding of the transcription factor ap-2. *Nucleic Acids Res*, 18(13):3975–82, 1990.

[22] Rene Cortese, Oliver Hartmann, Kurt Berlin, and Florian Eckhardt. Correlative gene expression and dna methylation profiling in lung development nominate new biomarkers in lung cancer. *The International Journal of Biochemistry & Cell Biology*, 40(8):1494 – 1508, 2008.

[23] G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, Y. Chen, J. A. Bernat, D. Ginsburg, D. Zhou, S. Luo, T. J. Vasicek, M. J. Daly, T. G. Wolfsberg, and F. S. Collins. Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). *Genome Res*, 16(1):123–31, 2006.

[24] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identificationid - 91, 2005.

[25] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. Pan, M. J. Schilstra, P. J. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295(5560):1669–78+, 2002.

[26] Florian Eckhardt, Joern Lewin, Rene Cortese, Vardhman K K. Rakyan, John Attwood, Matthias Burger, John Burton, Tony V V. Cox, Rob Davies, Thomas A A. Down, Carolina Haefliger, Roger Horton, Kevin Howe, David K K. Jackson, Jan Kunde, Christoph Koenig, Jennifer Liddle, David Niblett, Thomas Otto, Roger Pettett, Stefanie Seemann, Christian Thompson, Tony West, Jane Rogers, Alex Olek, Kurt Berlin, and Stephan Beck. Dna methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, October 2006.

[27] O. Elemento, N. Slonim, and S. Tavazoie. A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell*, 28(2):337–50+, 2007.

[28] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8+, 2007.

[29] P. C. FitzGerald, D. Sturgill, A. Shyakhtenko, B. Oliver, and C. Vinson. Comparative genomics of drosophila and human core promoters. *Genome Biol*, 7(7):R53+, 2006.

[30] Peter C. FitzGerald, Andrey Shlyakhtenko, Alain A. Mir, and Charles Vinson. Clustering of dna sequences in human promoters. *Genome Research*, 14(8):1562–1574, August 2004.

[31] Socorro Gama-Castro, Veronica Jimenez-Jacinto, Martin Peralta-Gil, Alberto Santos-Zavaleta, Monica I. Penaloza-Spinola, Bruno Contreras-Moreira, Juan Segura-Salazar, Luis Muniz-Rascado, Irma Martinez-Flores, Heladia Salgado, Cesar Bonavides-Martinez, Cei Abreu-Goodger, Carlos Rodriguez-Penagos, Juan Miranda-Rios, Enrique Morett, Enrique Merino, Araceli M. Huerta, Luis Trevino-Quintanilla, and Julio Collado-Vides. Regulondb (version 6.0): gene

regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucl. Acids Res.*, 36(suppl_1):D120–124, January 2008.

[32] M. Gardiner-Garden and M. Frommer. Cpg islands in vertebrate genomes. *J Mol Biol*, 196(2):261–82, 1987.

[33] K. Gaston and M. Fried. Cpg methylation and the binding of yy1 and ets proteins to the surf-1/surf-2 bidirectional promoter. *Gene*, 157(1-2):257–9, 1995.

[34] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–6+, 2002.

[35] M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, 2007.

[36] Y. Halperin, C. Linhart, I. Ulitsky, and R. Shamir. Allegro: analyzing expression and sequence in concert to discover regulatory programs. *Nucleic Acids Res*, 37(5):1566–79+, 2009.

[37] N. D. Heintzman and B. Ren. The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. *Cell Mol Life Sci*, 64(4):386–400, 2007.

[38] G. Z. Hertz and G. D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–77+, 1999.

[39] Da W. Huang, Brad T. Sherman, Qina Tan, Jack R. Collins, Gregory W. Alvord, Jean Roayaei, Robert Stephens, Michael W. Baseler, Clifford H. Lane, and Richard A. Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9):R183+, September 2007.

[40] Da W. Huang, Brad T. Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, Robert Stephens, Michael W. Baseler, H. Clifford Lane, and Richard A. Lempicki. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucl. Acids Res.*, 35(Web Server issue):gkm415+, June 2007.

[41] H. Ji, S. A. Vokes, and W. H. Wong. A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors. *Nucleic Acids Res*, 34(21):e146+, 2006.

[42] P. A. Jones and S. B. Baylin. The epigenomics of cancer. *Cell*, 128(4):683–92, 2007.

[43] T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–80, 2005.

[44] W. K. Kim, C. Krumpelman, and E. M. Marcotte. Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. *Genome Biol*, 9 Suppl 1:S5+, 2008.

[45] O. D. King, R. E. Foulger, S. S. Dwight, J. V. White, and F. P. Roth. Predicting gene function from patterns of annotation. *Genome research*, 13(5):896–904, May 2003.

[46] R. D. Kornberg and Y. Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98(3):285–94+, 1999.

[47] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14+, 1993.

[48] C. E. Lawrence and A. A. Reilly. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1):41–51+, 1990.

[49] R. D. Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Mol Syst Biol*, 4:213+, 2008.

[50] I. Lee, Z. Li, and E. M. Marcotte. An improved, bias-reduced probabilistic functional gene network of baker's yeast, saccharomyces cerevisiae. *PLoS ONE*, 2(10):e988+, 2007.

[51] M. D. Litt, M. Simpson, M. Gaszner, C. D. Allis, and G. Felsenfeld. Correlation between histone lysine methylation and developmental changes at the chicken beta-globin locus. *Science*, 293(5539):2453–5, 2001.

[52] X. Liu, D. L. Brutlag, and J. S. Liu. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, pages 127–38+, 2001.

[53] K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–60+, 1997.

[54] Kenzie D. MacIsaac and Ernest Fraenkel. Practical strategies for discovering regulatory dna sequence motifs. *PLoS Comput Biol*, 2(4):e36+, April 2006.

[55] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7+, 2006.

[56] L. Marino-Ramirez, J. L. Spouge, G. C. Kanga, and D. Landsman. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res*, 32(3):949–58, 2004.

[57] G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 2006.

[58] Alexander Meissner, Tarjei S. Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E. Bernstein, Chad Nusbaum, David B. Jaffe, Andreas Gnirke, Rudolf Jaenisch, and Eric S. Lander. Genome-scale dna methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770, August 2008.

[59] S. Myhre, H. Tveit, T. Mollestad, and A. Laegreid. Additional gene ontology structure for improved biological reasoning. *Bioinformatics*, 22(16):2020–2027, August 2006.

[60] M. E. Newman. Analysis of weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(5 Pt 2):056131+, 2004.

[61] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2 Pt 2):026113+, 2004.

[62] K. Noma, C. D. Allis, and S. I. Grewal. Transitions in distinct histone h3 methylation patterns at the heterochromatin domain boundaries. *Science*, 293(5532):1150–5, 2001.

[63] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–8+, 2005.

[64] Giulio Pavesi, Paolo Mereghetti, Giancarlo Mauri, and Graziano Pesole. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucl. Acids Res.*, 32(suppl_2):W199–203, July 2004.

[65] G. C. Prendergast and E. B. Ziff. Methylation-sensitive sequence-specific dna binding by the c-myc basic region. *Science*, 251(4990):186–9, 1991.

[66] M. Ptashne and A. Gann. Transcriptional activation by recruitment. *Nature*, 386(6625):569–77, 1997.

[67] J. W. Puckett, K. A. Muzikar, J. Tietjen, C. L. Warren, A. Z. Ansari, and P. B. Dervan. Quantitative microarray profiling of dna-binding molecules. *J Am Chem Soc*, 129(40):12310–9, 2007.

[68] Elizabeth Rach, Hsiang Y. Yuan, William Majoros, Pavel Tomancak, and Uwe Ohler. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the drosophila genome. *Genome Biology*, 10(7):R73+, 2009.

[69] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–9, 2000.

[70] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nat Biotechnol*, 16(10):939–45+, 1998.

[71] Julian M. Rozenberg, Andrey Shlyakhtenko, Kimberly Glass, Vikas Rishi, Maxim V. Myakishev, Peter C. Fitzgerald, and Charles Vinson. All and only cpg containing sequences are enriched in promoters abundantly bound by rna polymerase ii in multiple tissues. *BMC Genomics*, 9:67+, February 2008.

[72] P. J. Sabo, R. Humbert, M. Hawrylycz, J. C. Wallace, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos. Genome-wide identification of dnasei hypersensitive sites using active chromatin sequence libraries. *Proc Natl Acad Sci U S A*, 101(13):4537–42, 2004.

[73] D. Schubeler, D. M. MacAlpine, D. Scalzo, C. Wirbelauer, C. Kooperberg, F. van Leeuwen, D. E. Gottschling, L. P. O'Neill, B. M. Turner, J. Delrow, S. P. Bell, and M. Groudine. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev*, 18(11):1263–71, 2004.

[74] Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31(1):64–68, April 2002.

[75] S. Sinha and M. Tompa. Ymf: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, 31(13):3586–8+, 2003.

[76] S. T. Smale and J. T. Kadonaga. The rna polymerase ii core promoter. *Annu Rev Biochem*, 72:449–79, 2003.

[77] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan L. Rector, and Cornelius Rosse. Relations in biomedical ontologies. *Genome biology*, 6(5):R46+, 2005.

[78] Robert Stevens, Carole A. Goble, and Sean Bechhofer. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform*, 1(4):398–414, January 2000.

[79] G. D. Stormo and Gw Hartzell. Identifying protein-binding sites from unaligned dna fragments. *Proc Natl Acad Sci U S A*, 86(4):1183–7+, 1989.

[80] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[81] M. N. Swartz, T. A. Trautner, and A. Kornberg. Enzymatic synthesis of deoxyribonucleic acid. xi. further studies on nearest neighbor base sequences in deoxyribonucleic acids. *J Biol Chem*, 237:1961–7, 1962.

[82] D. Takai and P. A. Jones. Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A*, 99(6):3740–3745, March 2002.

[83] K. Tharakaraman, O. Bodenreider, D. Landsman, J. L. Spouge, and L. Mariño Ramírez. The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. *Nucleic Acids Res*, 36(8):2777–86+, 2008.

[84] The_gene_ontology_consortium. Creating the gene ontology resource: Design and implementation. *Genome Res.*, 11(8):1425–1433, August 2001.

[85] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, June 1998.

[86] M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schübeler. Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells. *Nat Genet*, 37(8):853–62+, 2005.

[87] Michael Weber, Ines Hellmann, Michael B. Stadler, Liliana Ramos, Svante Paabo, Michael Rebhan, and Dirk Schubeler. Distribution, silencing potential and evolutionary impact of promoter dna methylation in the human genome. *Nature Genetics*, 39(4):457–466, March 2007.

[88] F. Weih, D. Nitsch, A. Reik, G. Schutz, and P. B. Becker. Analysis of cpg methylation and genomic footprinting at the tyrosine aminotransferase gene: Dna methylation alone is not sufficient to prevent protein binding in vivo. *Embo J*, 10(9):2559–67, 1991.

[89] A. S. Weinmann and P. J. Farnham. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods*, 26(1):37–47, 2002.

[90] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. *Nature*, 434(7031):338–45, 2005.

[91] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*, 101(16):5934–9+, 2004.

[92] Muhammed A. Yildirim, Kwang-Il Goh, Michael E. Cusick, Albert-Laszlo Barabasi, and Marc Vidal. Drugtarget network. *Nature Biotechnology*, 25(10):1119–1126, October 2007.

[93] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabási, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–10+, 2008.

[94] J. Zhu, P. Y. Lum, J. Lamb, D. GuhaThakurta, S. W. Edwards, R. Thieringer, J. P. Berger, M. S. Wu, J. Thompson, A. B. Sachs, and E. E. Schadt. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenetic and genome research*, 105(2-4):363–374, 2004.